Extraction et analyse de données sur la gestion des adventices dans les systèmes de culture tropicaux

Université Grenoble-Alpes
M2 MIASHS SSD
Année universitaire 2019 - 2020

Auteur
Benjamin Fayolle

Tuteurs
Sandrine Auzoux
Thomas Le Bourgeois
Pascal Marnotte

Enseignant référent
Adeline Leclercq Samson

Table des Matières

1 Préambule

Ce rapport détaille le travail effectué de mars à septembre 2020 dans le cadre de mon stage de seconde année de master MIASHS, parcours Statistiques et Sciences des Données. Ce stage a été réalisé au sein du CIRAD, l'organisme français de recherche agronomique et de coopération internationale pour le développement durable des régions tropicales et méditerranéennes, et concerne un travail sur des données portant sur les mauvaises herbes qui se développent dans les diverses parcelles cultivées en régions tropicales. J'ai effectué ce stage au sein de l'unité AIDA (Agroécologie et Intensification Durable des Cultures Annuelles), et en partenariat avec l'unité AMAP (botAnique et Modélisation de l'Architecture des Plantes et des végétations), deux des 33 unités de recherches du CIRAD. Durant ces six mois, j'ai été encadré par Sandrine Auzoux[1], cadre informatique scientifique, Pascal Marnotte[1] et Thomas Le Bourgeois[2] , tous deux chercheurs en malherbologie. L'objectif principal de ce stage était de capitaliser une vaste quantité de données concernant les adventices tropicales, récoltées par divers malherbologues et agronomes au fil des années, dans des cultures tropicales.

En conséquence du contexte sanitaire de cette année 2020, le stage s'est déroulé dans des conditions particulières. Au lieu d'effectuer l'intégralité de celui-ci sur le site du CIRAD à Saint-Denis de La Réunion comme il était initialement prévu, la majorité du travail a été effectué à distance : après deux semaines sur site, j'ai travaillé depuis mon domicile, à Sainte-Clotilde sur l'île de La Réunion, jusqu'à la fin du mois de juin, avant de rentrer en métropole et d'effectuer les deux mois restants depuis Grenoble. Dans ce contexte, je tiens à remercier sincèrement Pascal, Thomas et Sandrine pour tout ce qu'ils ont fait pour moi durant cette période. Vous remercier d'abord de m'avoir permis de faire ce stage, mais surtout d'avoir toujours été réactifs, et de vous être continuellement assurés que tout allait bien pour moi qui était confiné à presque 9000 km de chez moi. Bien que je n'ai pas profité de l'île intense autant que je ne l'aurai souhaité, j'ai indéniablement beaucoup appris durant ces six mois, et travailler avec vous fut un réel plaisir.

[1]CIRAD, UR AIDA
[2]CIRAD, UR AMAP

2 Introduction

Les adventices des cultures représentent la principale cause de perte de récolte, tout particulièrement dans les régions tropicales où la température et l'humidité leur permettent de se développer et de croître rapidement. Dans ces régions, leur nocivité potentielle est beaucoup plus élevée que dans les conditions tempérées, et le temps consacré à la gestion de l'enherbement dans les systèmes de culture représente jusqu'à 50% du temps total consacré à une culture, depuis la préparation du sol jusqu'à la récolte. Par ailleurs, en cultures vivrières, la surface cultivable par une exploitation dépend directement de la capacité à gérer l'enherbement : un agriculteur ne peut en effet cultiver que la surface qu'il est en mesure de désherber. Or, dans ces zones tropicales, différentes formes d'agriculture coexistent, de l'agriculture traditionnelle avec très peu de moyens techniques et économiques, à des systèmes d'agriculture intensive, fondés sur des intrants chimiques et de la mécanisation. Il en résulte une diversité de pratiques de gestion des agroécosystèmes, liées à la diversité des cultures, aux conditions climatiques et pédologiques, et aux contextes socio-économiques rencontrées, ce qui se traduit en fin de compte par une diversité de communautés de mauvaises herbes et de contraintes liées à ces dernières.

L'étude du comportement des enherbements dans les systèmes de cultures tropicaux permet d'apporter des connaissances nécessaires à la mise en oeuvre de pratiques de gestion plus efficaces. Dans cette optique, plusieurs malherbologues liés au CIRAD se sont attelés à des études sur le terrain, aussi bien dans des parcelles expérimentales que sur des parcelles d'agriculteurs, afin d'étudier les plantes spontannées nuisibles aux cultures. Ainsi, bon nombre de relevés floristiques (inventaire des espèces présentes sur une parcelle, et dans certain cas d'indices de leur nuisibilité) ont été conduit sur des parcelles situées dans différents pays, concernant différentes cultures ou différents types de climats dits tropicaux. Toutes ces études ont été menées, depuis 30 ans, pour répondre à une question précise, dans un contexte particulier, et à en cela permis d'améliorer la connaissance locale des mauvaises herbes dans les cultures tropicales.

Toutefois, les données récoltées via ces relevés n'ont jamais été utilisées que localement : aucune étude n'a été menée sur l'ensemble des données, avec un objectif de compréhension du comportement des communautés de mauvaises herbes à une échelle globale. Les raisons à cela ne manquent pas. D'une part, beaucoup de ces études ont été menées par différents chercheurs, pas toujours en contact, mais surtout à des époques et des endroits différents. D'autre part, à l'image de la diversité des mauvaises herbes en cultures tropicales, ces études répondaient à une diversité d'objectifs différents, et il manquait en ce sens d'un format homogène permettant des analyses globales à grande échelle.

C'est dans ce contexte que trois chercheurs au CIRAD, Pascal Marnotte[1], Sandrine Auzoux[1], et Thomas Le Bourgeois[2], ont cherché à valoriser cette masse de données en l'utilisant pour analyser la répartition et la nuisibilité des adventices tropicales de manière globale. Pour cela, il est nécessaire de pouvoir aborder cette analyse à l'échelle de la parcelle, de l'exploitation, de l'ensemble d'un système de culture ou encore à l'ensemble des régions tropicales. Il est ainsi nécessaire de rassembler et d'associer tous les jeux de données préalablement collectés, par eux-même ou par d'autres auteurs, avec leur accord, et d'homogénéiser leur format afin de rendre possibles des analyses transversales à plusieurs jeux de données. Il est tout aussi indispensable de rendre accessible pour d'autres études cette masse de données, ainsi que de développer une chaine d'analyse de données cohérente et fonctionnelle, facilement réutilisable sur différents jeux de données.

Dans ce contexte, le présent travail a plusieurs objectifs : 1) rassembler, nettoyer, homogénéiser et publier d'anciens ensembles de données sur les mauvaises herbes tropicales, afin de les rendre librement disponibles pour des études plus approfondies, 2) mettre en place un système de concaténation et de filtres sur ces jeux de données dans le but de réaliser des analyses à un niveau de finesse contrôlé, 3) penser et implémenter une chaîne complète de traitement et d'analyse des données.

Ce rapport détaille ainsi les choix effectués et la méthodologie utilisée pour répondre à ces objectifs plus que les résultats proprements dits, dans la mesure où la quantité de données est vouée à augmenter à mesure que de nouveaux relevés seront effectués.

3 Extraction et traitement des données

Dans un premier temps, il a été nécessaire de rassembler et de mettre en forme autant de jeux de données que possible, en vue des futures analyses. En pratique, cette première étape peut elle-même être scindée en trois tâches distinctes. La première concerne la collecte des données. La seconde consiste en la standardisation de tous les jeux de données à disposition, afin de faciliter le traitement de ceux-ci. Enfin, une fois les données uniformisées, nous nous sommes attachés à les rendre disponibles via une plateforme de stockage publique de données.

3.1 Collecte des données

Première brique de tout travail d'analyse de données, la collecte de celles-ci a dans notre cas débuté en amont de ce stage. Il convient toutefois de distinguer dès à présent deux éléments : la collecte des données originelle, menée par différents chercheurs, doctorants, ou stagiaires, directement sur le terrain, et ayant donné lieu à la création de différents jeux de données, et la collecte de ces différents jeux de données auprès de leur(s) détenteurs(s) légitime(s), si ce(s) dernier(s) étai(en)t intéressé(s) par le projet, afin de les rassembler en une seule étude plus générale. Le premier de ces deux types de collecte est un travail de longue haleine, indépendant de ce stage, ayant débuté il y a quelques mois à plusieurs décennies selon l'étude. Le second type de collecte nous concerne directement, puisqu'il a constitué une première étape indispensable à ce stage : à l'heure actuelle, nous avons ainsi récolté 25 jeux de données de 14 auteurs les ayant eux-même récoltés directement sur le terrain, entre 1990 et 2020, dans des cultures tropicales.

Ces jeux de données originels sont tous des relevés floristiques, c'est-à-dire un inventaire des espèces végétales spontannée sur une parcelle agricole donnée. Toutefois, leur format varie selon les études, et sont au nombre de 6 :

- Relevés phytoécologiques non-pondérés, dans lesquels sont listées les espèces présentes, indépendamment de leur abondance respectives (relevés en présence/absence).
- Relevés phytoécologiques pondérés, dans lesquels les espèces présentes sont recensées et notées selon une échelle d'abondance spécifique (échelle 1-5 Braun-Blanquet [5], échelle 1-9 CEB [35, 36] ou pourcentage de recouvrement).
- Relevés sur des parcelles expérimentales avec mesures répétées dans le temps sur différentes parcelles.
- Relevés d'espèces dominantes : seules les espèces les plus abondantes ont été suivies.
- Relevés de synthèse : liste des espèces adventices par culture, zone ou pays.
- Synthèse des parcelles témoins dans des essais expérimentaux sur des herbicides : score moyen d'abondance (échelle 1-9 CEB) sur les parcelles témoins (non-traitées).

Dans ces relevés, les espèces étaient nommées selon des nomenclatures variables d'un jeu de données à l'autre. Par ailleurs, à chaque relevé floristique était associé un certain nombre de facteurs environnementaux, dont la nature dépendait des besoins de l'étude originelle. Ces facteurs sont donc eux-aussi non standardisés, varient en niveau de détail autant qu'en quantité (entre 5 et 19 facteurs selon l'étude). Enfin, les jeux de données ont été rassemblés à partir de fichiers Excel, ou bien extraits de bases de données Access et convertis en fichiers Excel.

3.2 Standardisation et homogénéisation

Il est aisé de se rendre compte, lors de la collecte des jeux de données, qu'il n'existe aucune harmonisation entre ces derniers. Chaque jeu de données a en effet été créé pour répondre à une question spécifique, à une

période et dans un pays différent, par un auteur différent. Il est dès lors impossible de mettre en place des analyses systématiques sur ces jeux de données disparates. Au delà de l'analyse globale de l'ensemble de ces données, ce projet a pour objectif de capitaliser sur une vaste quantité de données récoltées et donc d'inciter les malherbologues des régions tropicales à partager leurs jeux de données, ainsi qu'à utiliser à leur guise les données déjà récoltées. Ce projet a donc une portée publique, et en cela, les jeux de données se doivent de partager une structure commune. Le second temps de ce travail sur les données a donc été un travail de standardisation et d'homogénéisation des jeux de données.

La première transformation réalisée concerne le format et le titre de chacun des jeux de données. Chaque fichier a vu son format être fixé sur un fichier Excel, version 2013, si ça n'était pas déjà le cas. Ensuite, une homogénéisation des titres des jeux de données a été mise en œuvre. Cette dernière est nécessaire, puisque le nombre de jeux de données est déjà élevé (25), et est voué à augmenter. Il faut donc que le titre permette d'avoir immédiatement accès à des informations de premier plan, telles que l'auteur ou l'année, mais aussi le pays, la culture concernée ou le type de notation utilisé dans le relevé floristique. Ainsi, l'homogénéisation des titres a été effectuée de la manière suivante : **3 lettres pour le pays - 3 lettres pour l'auteur - année - 3 lettres pour la culture - 2 lettres pour le type de notation**. Ainsi, et en guise d'exemple, le fichier **CAM-LEB-1990-DIV-AD.xlsx** correspond à l'étude menée par T. Le Bourgeois au Cameroun en 1990, sur diverses cultures, utilisant une notation floristique en abondance (échelle 1-9 CEB). La table présentée en Annexe 1.1 résume les différentes abréviations actuellement utilisées.

Une fois le format et les titres uniformisés, nous avons réorganisé et complété le contenu même des fichiers. Pour être plus précis, nous avons, pour chaque étude, divisé le fichier originel en plusieurs onglets, au sein du même fichier Excel. Les données originelles, telles que récoltées par l'auteur, ont été conservées dans un onglet, et trois nouveaux onglets ont été créés. Un premier recense un certain nombre de méta-données telles que l'objectif premier de l'étude et les détails de la transformation des notations originelles de la flore (voir Annexe 1.2). Un second contient tous les facteurs agro-environnementaux associés à l'étude originelle, tels que décrits par l'auteur, ainsi que de nouveaux facteurs (non renseignés mais facilement accessibles), si besoin, afin d'avoir un ensemble de facteurs communs à toutes les études. Ces facteurs communs sont l'auteur, l'année, le pays, la culture principale de la parcelle ainsi que le type général de culture, le climat et le type d'irrigation (inondé ou non). Le troisième onglet créé contient les notations floristiques transformées de l'étude. Notons que nous avons également pris soin de standardiser les formats des onglets facteurs et floristique (les espèces adventices ou les facteurs sont toujours en ligne, et les relevés toujours en colonne, afin de respecter le format courant des relevés en malherbologie), ainsi que les noms de relevés. Pour ces derniers, un nom générique composé de **3 lettres pour le pays - 3 lettres pour l'auteur - 3 chiffres pour le numéro de relevé** (par exemple, **CAM-LEB-001** pour le premier relevé de l'étude correspondante) a été fixé, afin de faciliter la lecture en cas de concaténation de plusieurs études.

La nomenclature des espèces adventices a ensuite été actualisée, afin de suivre les standards de *Plants Of The World Online* [37] et *Catalogue Of Life* [43]. Une fois ce travail de vérification effectué, les noms scientifiques des espèces ont été remplacés par leur code EPPO [10], qui code en 5 lettres le nom scientifique d'une plante. Pour les espèces n'en disposant pas déjà, de nouveaux codes ont été créés, marqués en rouge, et soumis à l'EPPO pour confirmation. Dès lors, un nouveau fichier Excel de référence concernant la nomenclature des espèces a été créé. Ce fichier contient les codes EPPO de toutes les espèces recensées au fil des différents jeux de données (soit actuellement 1530 espèces), ainsi que le nom scientifique, le nom de famille de l'espèce, et occasionnellement les synonymes les plus courants ou utilisés dans les études. L'objectif de la création de ce fichier est double : d'une part, en réunissant toutes les espèces au sein d'un même fichier, les mises à jour ultérieures de nomenclature sont facilitées. Les noms scientifiques des plantes sont en effet régulièrement révisés, alors que les codes EPPO ne changent pas. D'autre part, un code de 5 lettres sera bien plus lisible lors de futures représentations graphiques des résultats d'analyses.

Enfin, l'une des modifications les plus importantes apportées aux données originelles est la standardisation de la notation floristique utilisé dans les relevés. Comme nous l'avons vu plus haut, l'échelle de notation de l'abondance est différente selon les auteurs, et certaines études sont effectuées uniquement en présence/absence. Il a ainsi été décidé de garder comme échelle standard (pour les études en abondance) l'échelle 1-9 CEB. L'annexe 1.2 détaille les différentes transformations depuis les échelles originelles vers l'échelle standard. Encore une fois, l'objectif de cette action est de pouvoir analyser plusieurs études en même temps. De plus,

pour toutes les études en abondance, une version en présence/absence de l'onglet floristique a été générée. Ceci fait, les différents onglets ont été exportés au format texte en vue du stockage, de la mise à disposition et des analyses, étant donné qu'il s'agit probablement du format le plus général possible de données.

3.3 Stockage et mise à disposition

Les jeux de données ainsi retravaillés ont ensuite été stockés et publiés sur la plateforme Dataverse du CIRAD. Dataverse [8], projet porté et créé par l'*Institute for Quantitative Social Science* (IQSS) de l'université d'Harvard, est une plateforme digitale *open source* dont l'objectif principal est le stockage et le partage des données. A chaque jeu de données créé sur la plateforme est associé un certain nombre de méta-données précises, permettant la recherche de jeux de données par mots-clés, mais aussi précisant les conditions d'utilisation de ces derniers. De plus, chaque jeu de données publié reçoit un identifiant DOI (*Digital Object Identifier*), et peut ainsi être cité d'une manière unique incluant le nom de tous les auteurs, en cas d'utilisation externe. Les données ont été publiées sur un dataverse propre au CIRAD[3], sous la licence *Creative Commons Attribution 4.0 International* [1]. Le stockage des données a été organisé comme suit. Un sous-dataverse dédié spécifiquement aux études sur les adventices des cultures tropicales, nommé *Amatrop: Tropical Weed Studies*, a d'abord été créé[4] et lié au portail collaboratif Wiktrop (*Weed identification and Knowledge in the tropical and mediterranean areas*) [22–24]. À l'intérieur de ce sous-dataverse ont été créés autant de *datasets* que d'études originelles (soit, à l'heure actuelle, 25), plus un dataset contenant le fichier de référence de nomenclature des espèces. Chaque *dataset* (hormis le dernier) contient :

- Le fichier global, au format Excel 2013
- Un fichier texte correspondant à l'onglet métadonnées
- Un fichier texte correspondant à l'onglet facteurs
- Un fichier texte correspondant à l'onglet floristique, en présence/absence
- Un fichier texte correspondant à l'onglet floristique, en abondance, pour les études concernées

Une telle organisation permet de donner une certaine "autonomie" à chaque étude : chaque étude peut, indépendamment des autres, être consultée, utilisée, et citée. De plus, en procédant ainsi, il est possible de renseigner des métadonnées plus précises pour chaque jeu de données, et donc une recherche plus fine peut être effectuée au sein du sous-dataverse (par exemple, récupérer uniquement les jeux de données concernant tel pays, climat ou culture..).

3.4 Description des jeux de données

Le dataverse créé contient à l'heure actuelle 25 jeux de données, plus un fichier de référence recensant toutes les espèces adventices rencontrées dans les jeux de données, soit actuellement 1530 taxons différents. Parmi ceux-ci, 183 ne disposaient pas encore de code EPPO. Nous avons donc créé un code temporaire, et l'avons envoyé à l'EPPO pour confirmation. Les 25 études prises ensembles contiennent au total 7251 relevés floristiques. La taille des études varie fortement : l'étude la plus "petite" ne contient que 11 relevés floristiques, quand la plus "grande" en contient plus de 1200. La table 1 décrit le nombre d'espèces distinctes, de relevés, et d'enregistrements total d'espèces au sein de chaque jeu de données.

Les 25 jeux de données actuellement concernés par l'étude ont été récoltés dans 9 pays tropicaux différents, majoritairement en Afrique, mais aussi dans une moindre mesure en Asie et en Amérique du Sud. Ils concernent 6 climats tropicaux différents, ainsi que 11 types de cultures. Notons tout de suite qu'un jeu de données ne concerne qu'un pays, mais peut contenir des données suffisamment espacées géographiquement

[3] https://dataverse.cirad.fr/
[4] https://dataverse.cirad.fr/dataverse/amatrop

Table 1: Quantification des données par étude

Etude	**Nombre de relevés**	**Nombre d'espèces distinctes**	**Nombre d'enregistrements**
BEN-MAR-2013-DIV-AD	26	86	228
CAM-LEB-1988-DIV-AD	178	206	3 657
CAM-MAR-1999-MSK-AD	52	64	573
CAM-VAL-2001-COT-AD	1 217	71	11 653
CDI-AWA-2011-VIV-AD	140	389	8 111
CDI-BOR-1997-CAN-PA	544	231	7 878
CDI-GNA-1997-JAC-AD	64	54	1 303
CDI-IPO-2009-DIV-PA	11	494	964
CDI-MAR-1991-CAN-AD	261	124	2 094
CDI-SAB-2015-VIV-AD	200	324	7 922
CDI-TEH-2014-COT-AD	619	66	1 835
GUI-MAR-1996-COT-AD	110	158	1 632
GUY-LEB-2018-DIV-AD	61	137	735
MAD-ANT-2017-RIZ-AD	1 047	81	4 131
MAD-JAR-1998-DIV-AD	198	120	3 193
MAD-JA2-1999-DIV-AD	210	140	3 839
MAD-RAK-2015-RIZ-AD	544	79	4 983
MAU-MAR-2019-MAR-AD	82	206	1 316
RUN-BAI-2017-CAN-AD	219	73	2 978
RUN-LEB-2003-DIV-AD	566	278	13 420
RUN-MAR-2016-CAN-AD	108	17	1 666
RUN-MA2-2019-CAN-AD	121	62	1 615
RUN-MA3-2019-CAN-AD	170	39	1 606
RUN-VIA-2019-CAN-AD	40	54	462
VIE-STE-1999-RIZ-AD	363	113	5 045
Total	7 251	1 530	92 839

pour couvrir plusieurs climats. Il en va de même pour les cultures : une étude peut contenir plusieurs types de cultures. La Figure 1 résume le nombre d'études concernées par pays, cultures et climats. Par ailleurs, 13 des 25 études actuelles sont multi-sites, c'est-à-dire qu'elle ont été menées sur différents sites dans un pays. Les 12 autres sont mono-site. Cette information est disponible dans le tableau présenté en Annexe 1.3.

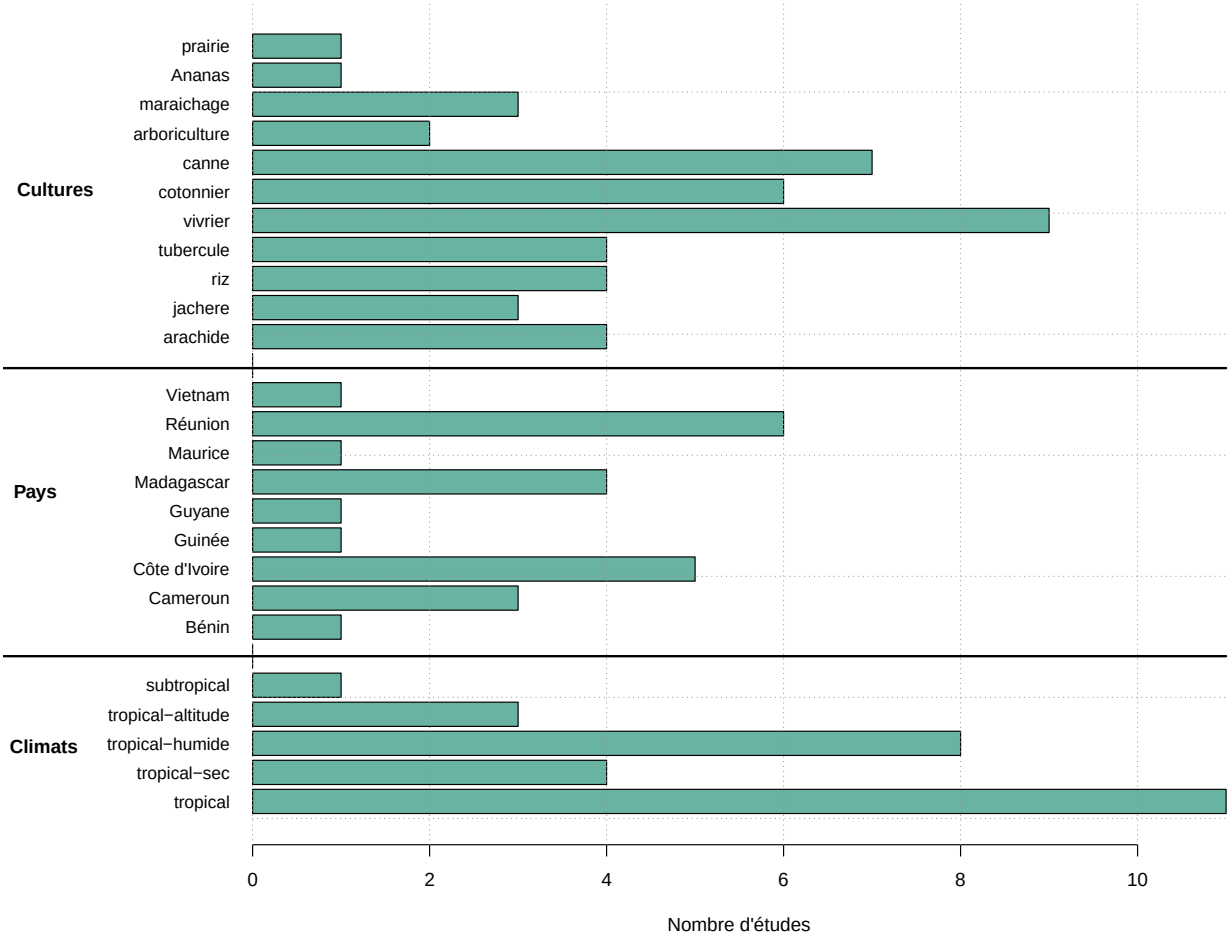

Figure 1: **Nombre d'études concernées par pays, climats et cultures**.

4 Analyses statistiques simples

Nous avons jusqu'ici réalisé une première étape de traitement des données. Nous avons rassemblé, standardisé et mis à disposition sur la plateforme publique Dataverse 25 jeux de données. Cette capitalisation sur des données portant sur les adventices tropicales, récoltées à travers le monde depuis plusieurs décennies, constitue en soi la réalisation d'un premier objectif de ce projet. Cette première étape accomplie, le second temps de ce stage consiste à profiter du format désormais homogène des jeux de données pour mettre en œuvre des analyses systématiques, à différents niveaux, afin de caractériser le comportement des espèces adventices dans les cultures tropicales à une échelle globale.

4.1 Diagrammes d'infestation

Nous avons donc décidé de mettre en œuvre, dans un premier temps, une série d'analyses portant sur les onglets floristiques des jeux de données standardisés. L'objectif principal de cette première série d'analyses est d'avoir une vue globale de la fréquence et de l'abondance des espèces au sein des cultures tropicales, afin de repérer les espèces particulièrement fréquentes et/ou abondantes, et donc dommageables pour les cultures.

Afin d'atteindre ces objectifs, nous avons développé différentes chaînes de traitement et d'analyse des données. Dans un premier temps, nous avons croisé la fréquence relative avec l'abondance moyenne locale des espèces adventices, au sein de graphiques nommés diagrammes d'infestation [19]. L'adjectif 'locale' renvoie au fait que l'abondance moyenne des espèces est calculée uniquement sur les relevés dans lesquels l'espèce est présente. Pour cela, il est d'abord nécessaire d'exclure les études pour lesquelles on ne dispose que de relevés en présence/absence (soit à ce jour deux études : **CDI-BOR-1997-CAN-PA** et **CDI-IPO-2009-DIV-PA**). Ensuite, à partir de l'onglet floristique, on calcule la fréquence relative RF_s de chaque espèce s comme le nombre de fois où l'espèce apparaît dans un relevé divisé par le nombre total de relevés. En notant M la matrice d'abondance des espèces pour un ensemble de relevés floristiques, avec la cellule $M_{s,r}$ correspondant à la valeur d'abondance de l'espèce s dans le relevé r, et N_r le nombre total de relevés, on obtient donc :

$$RF_s = \frac{\sum_r \mathbb{1}_{(M_{s,r} \neq 0)}}{N_r} \tag{1}$$

On calcule de même l'abondance moyenne locale AM_s de chaque espèce, dans les relevés où elle est présente, comme :

$$AM_s = \frac{\sum_r M_{s,r}}{\sum_r \mathbb{1}_{(M_{s,r} \neq 0)}} \tag{2}$$

On peut dès lors analyser graphiquement la fréquence relative et l'abondance moyenne locale des espèces au sein d'un jeu de données en considérant ces deux quantités comme coordonnées. On obtient ainsi un diagramme d'infestation, dont un exemple est présenté en Figure 2, à partir duquel on peut mettre en évidence quatre types d'espèces en fonction de leur comportement :

- les espèces fréquentes et régulièrement abondantes, comme AGECO (*Ageratum conizoides*), que l'on considèrera comme des adventices "majeures générales", qui posent le plus de problèmes de gestion aux agriculteurs, dans l'étude considérée.
- les espèces fréquentes mais rarement abondantes, comme MAPFL (*Cyperus tenuis*) ou PYLAM (*Phyllanthus amarus*), que l'on considèrera comme des adventices générales, qui pour l'instant posent peu de problèmes de gestion malgré leur fréquence forte.

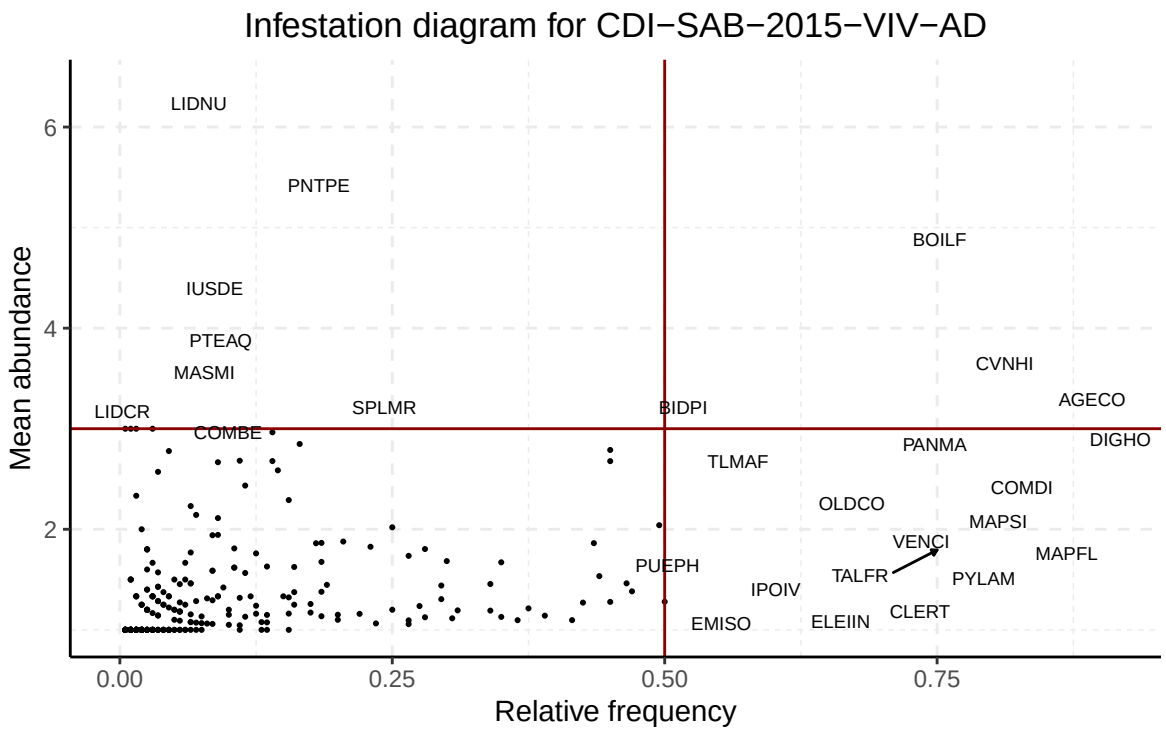

Figure 2: **Diagramme d'infestation pour l'étude CDI-SAB-2015-VIT-AD**. Par souci de lisibilité, seuls les codes EPPO correspondant aux espèces suffisamment fréquentes (seuil de 0.5) ou suffisamment abondantes (seuil de 3) ont été affichés. La partie haute droite regroupe les adventices majeures générales ; la partie basse droite correspond aux adventices générales, fréquentes mais peu abondantes ; la partie haute gauche contient les espèces majeures locales ; et enfin la partie basse gauche correspond aux espèces adventices mineures.

- les espèces peu fréquentes mais abondantes quand elles sont présentes, comme LIDNU (*Craterostigma nummulariifolium*) ou PNTPE (*Pentodon pentandrus*), que l'on considèrera comme des espèces "majeures locales" qui sont difficiles à gérer lorsqu'elles sont présentes. Ces espèces sont généralement inféodées à des milieux particuliers peu représentés dans la région d'étude.
- les espèces peu fréquentes et jamais abondantes qui correspondent aux espèces "mineures" qui ne présentent pas de problèmes de gestion ou de compétition avec les cultures.

4.2 Dépendance et corrélations entre espèces

L'analyse par les diagrammes d'infestation permet de faire ressortir, à partir de l'onglet floristique d'un jeu de données, les espèces fréquentes et/ou abondantes, et donc de repérer aisément les espèces adventices les plus nuisibles ou les plus difficiles à gérer dans les cultures concernées. Toutefois, les espèces sont ici considérées indépendamment les unes des autres. Or, un second objectif, complémentaire au premier, est de mieux cerner les liens existants entre différentes espèces : par exemple, est-ce que telle espèce est nécessairement accompagnée de telle autre espèce ? Au contraire, il peut être très utile de savoir qu'une espèce ne peut se développer en présence d'une autre. Nous avons ainsi cherché à mettre en évidence ces liens entre espèces, prises deux à deux. Nous nous sommes d'abord attachés à repérer les espèces présentant une dépendance statistique dans leur présence : à partir de l'onglet floristique, nous avons effectué un test exact de Fisher [11]

sur les tables de contingence de la présence/absence de toutes les espèces prises deux à deux. Nous avons préféré un test exact de Fisher à un test du Khi deux, testant lui aussi l'indépendance de deux variables, dans la mesure où ce dernier devient inexacte en cas de faible taille d'échantillons. Or, il est parfaitement envisageable de vouloir appliquer ces analyses à une culture particulière (par exemple), dans laquelle il y a peu de relevés. Le test exact de Fisher, lui, est comme son nom l'indique valide pour toutes les tailles d'échantillons. Nous avons ainsi calculé et représenté graphiquement la matrice des p-valeurs associées à ces tests. Par souci de lisibilité, nous nous sommes dans un premier temps limité aux 50 espèces les plus fréquentes pour la visualisation. Nous avons également choisi d'afficher uniquement les points pour lesquels la dépendance est significative au seuil de 5%, et nous avons représenté le $\log_{10}$ des p-valeurs. Ainsi, sur la Figure 3, qui montre un exemple de visualisation des dépendances, plus la valeur est faible, plus la dépendance est significative, et une absence de point correspond à une indépendance statistique entre deux espèces.

Cette dépendance est en outre à mettre en perspective de la co-occurrence des espèces. En effet, si le test de Fisher permet de repérer les espèces présentant une dépendance statistique dans leur présence/absence, il reste difficile de l'interpréter : les deux espèces sont-elles souvent présentes ensembles (dépendance positive), ou bien très rarement (dépendance négative, auquel cas on peut interpréter la dépendance comme : si l'espèce A est présente, alors l'espèce B a automatiquement peut de chances de l'être également) ? C'est pourquoi nous avons calculé et représenté une matrice de co-occurence entre espèces prises deux à deux, donnant la proportion de relevés dans lesquels deu espèces sont présentes ensembles. Bien que cette matrice présente ses défauts (les espèces généralement fréquentes auront plus de poids), elle a aussi l'intérêt de faciliter l'interprétation de la matrice de dépendance : si deux espèces sont statistiquement dépendantes, et co-occurrent beaucoup, alors la dépendance est problement positive, et inversement.

L'analyse des dépendances permet ainsi d'identifier des espèces liées statistiquement. Toutefois, elle ne tient aucun compte des valeurs d'abondance : si deux espèces dépendent l'une de l'autre, rien ne dit que leurs abondances aient des tendances comparables. Afin de prendre en considération ces valeurs d'abondance, le coefficient de corrélation de Spearman [45] a été calculé entre les relevés de toutes les espèces prises deux à deux. Si deux espèces dépendantes se développent de manière comparable en terme d'abondance, le coefficient de corrélation sera positif. À l'inverse, si, pour deux espèces se trouvant régulièrement ensemble, une occupe tout l'espace et étouffe l'autre, alors le coefficient de corrélation sera négatif. Bien sûr, des espèces co-occurrentes peuvent présenter des abondances non-corrélées. De même, un corrélation négative ne dénote pas nécessairement une compétition, mais potentiellement une adapatation à des milieux différents. À noter que le coefficient de Spearman a été préféré à celui de Pearson [12], dans la mesure où l'on s'intéresse à la relation monotonne entre les indices d'abondances de deux espèces plutôt qu'aux valeurs brutes et à une relation affine. Une matrice de corrélation entre les abondances des espèces est alors calculée, puis représentée graphiquement (là aussi, nous nous sommes d'abord limité aux 50 espèces les plus fréquentes, par souci de lisibilité). De plus, pour chaque corrélation calculée, un test d'hypothèse permettant de vérifier que la corrélation est significativement différente de 0 est mis en place. Finalement, seules sont gardées les corrélations pour lesquelles la p-valeur du test sous-jacent est inférieure à 5%. La Figure 4 donne un exemple de visualisation. Là encore, il est important de mettre en perspective cette matrice de corrélation avec celle des co-occurrences : il est en effet possible qu'une corrélation soit significative, alors même que les deux espèces sous-jacentes ne co-occurent que très peu.

4.3 Les différents axes d'analyse

Une fois les différents types d'analyses établis, nous avons décidé de les mettre en œuvre à différents niveaux. En effet, l'objectif général de cette étude est de concaténer des jeux de données variés pour une analyse globale des contraintes d'enherbement. Pour cela nous avons tiré profit de la base de facteurs communs entre tous les jeux de données afin de décider de filtres d'analyses pertinents. Ainsi, 4 niveaux d'analyses ont été choisis.

- Axe étude : les analyses pour chacune des études ont tout d'abord été gardées, dans la mesure où ces résultats pourraient intéresser l'auteur de l'étude originelle ayant abouti au jeu de données en question,

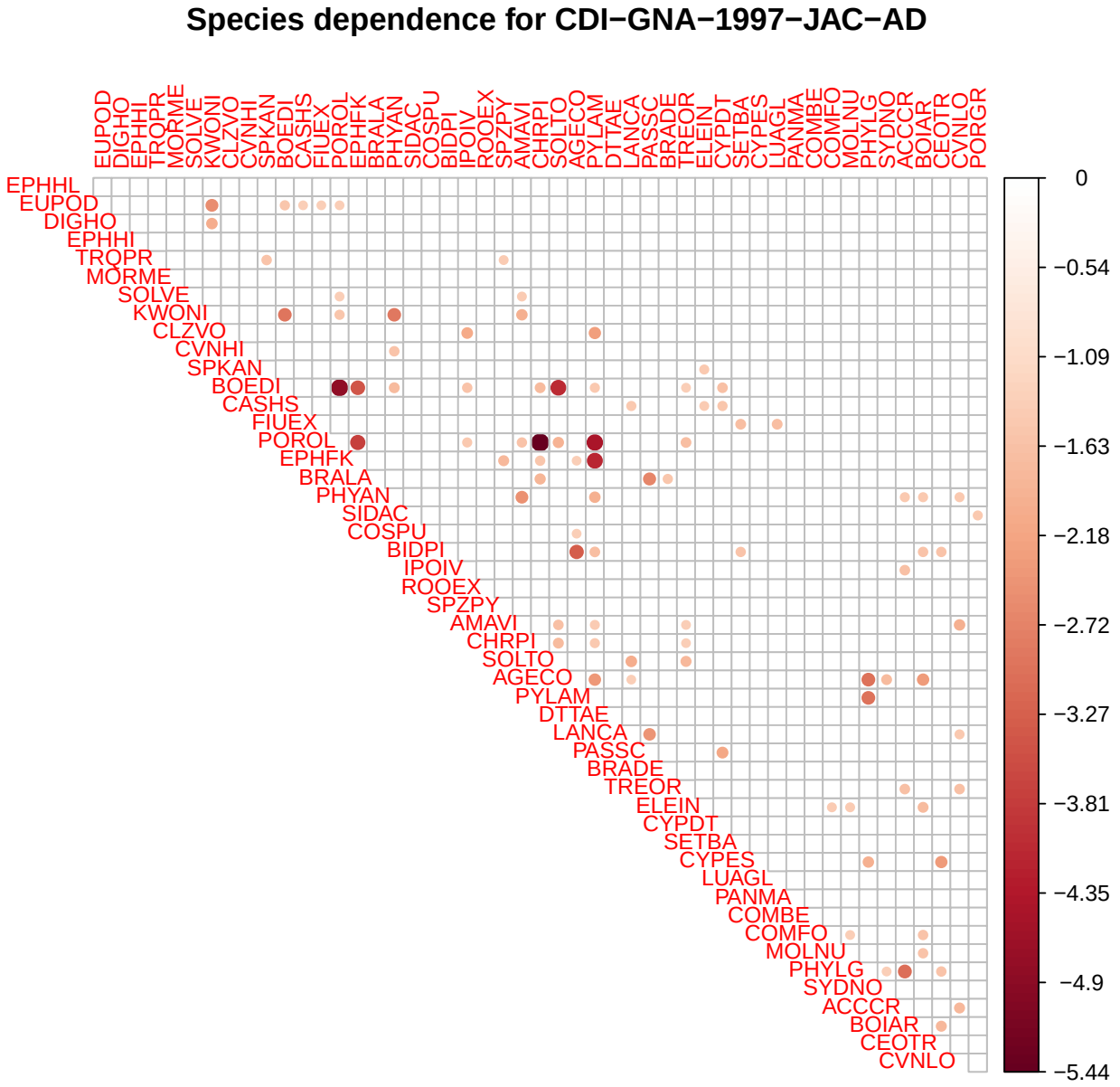

Figure 3: **Matrice de dépendance entre les espèces prises 2 à 2, pour l'étude CDI-GNA-1997-JAC-AD**. Les 50 espèces les plus fréquentes sont représentées. Le code couleur correspond au $\log_{10}$ de la p-valeur des tests exacts de Fisher.

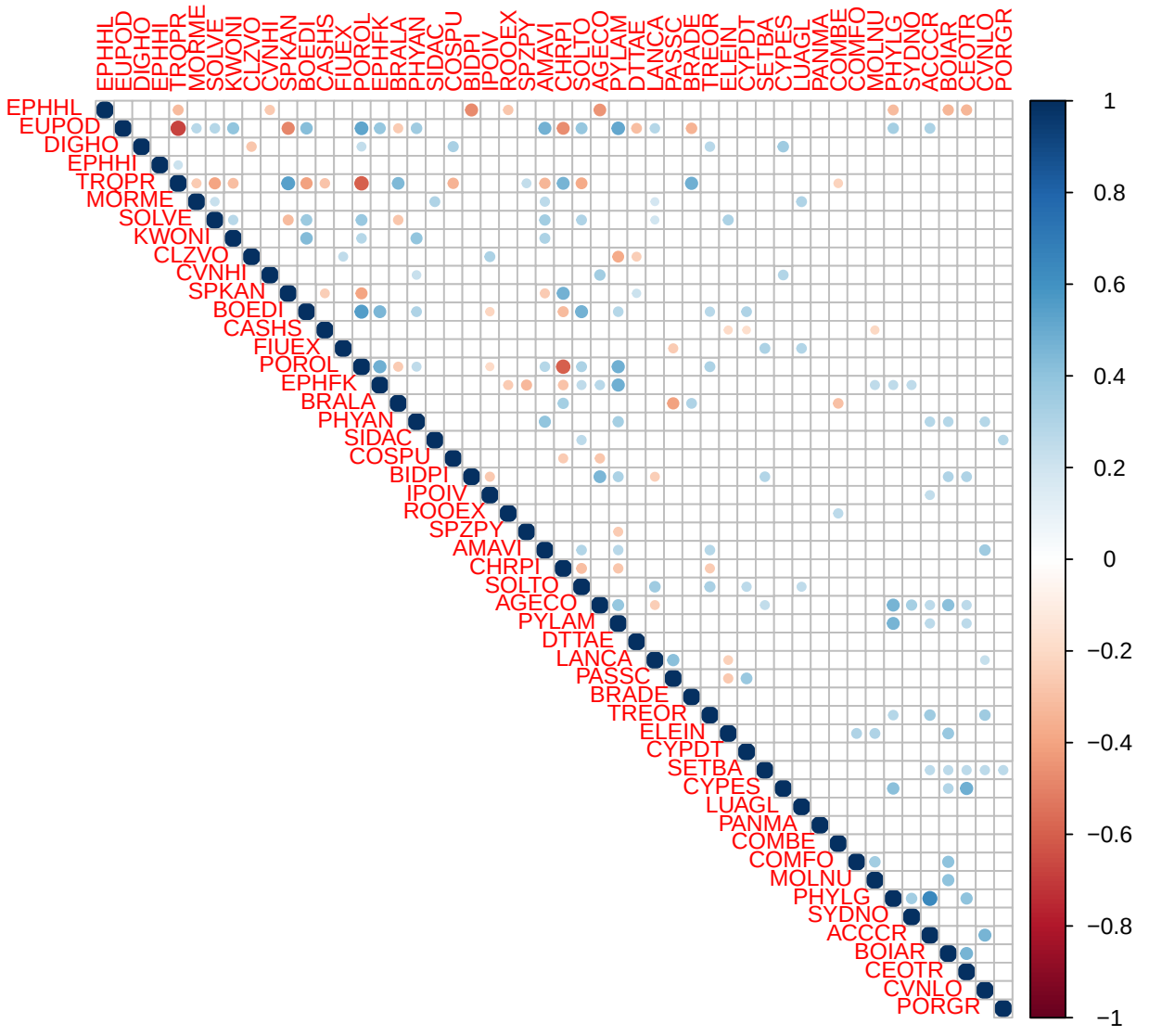

Figure 4: **Matrice de corrélation entre les espèces prises 2 à 2, pour l'étude CDI-GNA-1997-JAC-AD**. Les 50 espèces les plus fréquentes sont représentées.

ou même toute autre personne ayant travaillé plus ou moins étroitement avec l'étude originelle.

- Axe pays : les jeux de données ont été rassemblés et analysés par pays. À ce jour, 9 pays (ou régions, puisque des départements d'outre-mer français sont considérés comme distincts) sont concernés : le Bénin, le Cameroun, la Côte d'Ivoire, Madagascar, l'île de La Réunion, la Guyane française, la Guinée, l'île Maurice et le Viet Nam.
- Axe climat : les jeux de données ont été rassemblés et analysés par type de climat. À ce jour, 5 types de climats sont concernés : tropical, tropical-sec, tropical-humide, tropical-altitude et subtropical.
- Axe culture : une première distinction a d'abord été faite entre cultures pluviales et inondées, puisque la flore y est très différente. Ensuite, en comparant cultures pluviales et cultures inondées, une analyse globale, comprenant tous les relevés concernés, a été réalisée. Puis, les analyses ont été affinées par type de culture. À ce jour, deux cultures inondées (riz et vivrier) ainsi que 11 cultures pluviales (ananas, arachide, riz, vivrier, arboriculture, jachère, maraîchage, prairie, tubercule, cotonnier et canne à sucre) sont concernées.

Au-delà des analyses générales par axe, a été rendue possible la combinaison de ceux-ci par l'implémentation d'un système de filtres d'analyse. L'objectif est de laisser à l'utilisateur le choix d'un niveau de finesse voulu pour effectuer les analyses souhaitées. Ainsi, il est possible de combiner les axes présentés plus hauts de la manière voulue afin d'affiner l'analyse. De plus, pour un axe donné, le système de filtres créé permet de prendre en compte, ou d'exclure, chaque modalité de celui-ci. Il est donc possible de mettre en œuvre toutes les analyses simples présentées plus hauts pour comparer le cotonnier camerounais du cotonnier de Côte d'Ivoire, par exemple.

5 Analyses multivariées

La deuxième étape du stage a donc permis de tirer profit du format désormais homogène entre les différents jeux de données afin de mettre en place des analyses relativement simples dans la méthode, mais riches d'informations, et ce de manière systématique et selon différents axes d'analyse. À l'issue de celles-ci, il nous est possible de décrire la répartition des principales espèces adventices au sein des cultures tropicales, et nous avons également pu nous faire une première idée des relations existantes entre ces espèces. En revanche, ces analyses restent "simples", dans la mesure où elles portent uniquement sur l'abondance des espèces, sans mettre directement en relation celle-ci avec des facteurs environnementaux. Bien sûr, les différents axes d'analyses choisis permettent d'effectuer ces analyses en se cantonnant à une modalité précise d'un facteur (une culture, par exemple), mais cette relation n'est qu'indirecte. En effet, ce premier volet d'analyses ne permet pas de quantifier l'impact de potentiels facteurs environnementaux, et de mettre en évidence des "facteurs principaux" permettant d'expliquer l'abondance ou l'abondance de telle ou telle espèce. Un second volet d'analyses, multivariées, permet d'aborder cet aspect de l'analyse.

5.1 Analyse en Composantes Principales sur Variables Instrumentales (ACPVI)

L'une des raisons pour lesquelles nous n'avions jusqu'ici pas mis en œuvre d'analyses multivariées est le manque de facteurs environnementaux susceptibles d'expliquer les abondances d'espèces, qui soient communs à tous les jeux de données. En effet, outre la culture, le type d'irrigation et une description très générale du climat, la plupart des jeux de données ne disposent pas de facteurs environnementaux communs. Toutefois, plusieurs jeux de données relatifs à l'île de La Réunion font exception : en plus de la culture et du type d'irrigation, ces derniers disposent de facteurs environnementaux tels que l'altitude de la parcelle, la pluviométrie moyenne, mais aussi des informations relatives aux pratiques de désherbage et aux cultures précédentes. L'intérêt d'une analyse multivariée devient alors réel.

Le problème se posant alors est celui de la méthode. En effet, nous ne disposons pas d'une seule variable, que nous cherchons à expliquer par un ensemble de facteurs, mais d'un certain nombre d'espèces. En d'autres termes, nous disposons de deux tableaux de données, une matrice d'abondance des espèces et une matrice des descripteurs. Nous soupçonnons en outre que le second influence directement le premier. Or, la plupart des analyses multivariées classiques mettent en relation non pas deux tableaux, mais une variable dépendante et un tableau de variables explicatives.

Une manière potentielle de faire aurait été de réaliser et d'interpréter autant de régressions linéaires multiples qu'il y a d'espèces. Mais cette approche comporte nombre de points noirs : l'analyse n'est pas générale, elle se fait par espèce, rendant une interprétation globale difficile ; les facteurs environnementaux sont souvent liés entre eux (la culture est très corrélée à l'altitude, par exemple), rendant plus difficile encore l'interprétation, et allant à l'encontre des hypothèses de base d'un modèle de régression. L'Analyse Factorielle des Correspondances (AFC) permet de palier ces problèmes, en projetant l'ensemble des espèces et des descripteurs dans un même espace et en mettant en évidence les corrélations entre espèces et variables. Toutefois, l'AFC elle-même présente des inconvénients majeurs : en utilisant une table de contingence, on perd l'information de l'abondance, pourtant de premier ordre ; l'AFC n'admet pas de variable quantitative, et celles-ci doivent donc être converties en classes, ce qui induit nécessairement une perte d'information, ainsi que des seuils bien souvent arbitraires.

Ces limites inhérentes aux analyses multivariées classiques nous ont poussé à utiliser une autre méthode, moins répandue mais plus à même de répondre à nos besoins : l'Analyse en Composantes Principales sur Variables Instrumentales (ACPVI) [25, 44]. Cette méthode permet de mettre en relation nos deux tableaux de données. Introduisons dès maintenant les notations nécessaires :

- n : le nombre de relevés

- p : le nombre d'espèces présentes
- q : le nombre de descripteurs
- X : la matrice des descripteurs, de dimension (n, q)
- Y : la matrice d'abondance, de dimension (n, p)

L'ACPVI fonctionne en plusieurs étapes. La première consiste à réaliser p régressions linéaires multiples des abondances des espèces en fonction de la matrice X des descripteurs (les descripteurs qualitatifs seront rendus binaires selon les principes du *One Hot Encoding*). Les régressions se font dans une métrique D, c'est-à-dire que l'on pondère chaque observation par un poids. Dans la formulation classique de l'ACPVI, $D = \frac{1}{n} I_n$ (où I_n est la matrice identité d'ordre n). On exprime ainsi chaque espèce Y_i comme une combinaison linéaire des descripteurs, plus un résidu indépendant de ces derniers :

$$Y_i = \beta_{i,0} + \sum_{k=1}^{p} \beta_{i,k} X_k + \varepsilon_i,$$

c'est-à-dire :

$$Y_i = \widehat{Y_i} + \varepsilon_i \tag{3}$$

On peut ainsi constituer une nouvelle matrice $\widehat{Y}$ composée des estimation des abondances par régressions linéaires sur les descripteurs. Ainsi, on peut interpréter $\widehat{Y}$ comme la part des abondances strictement expliquée par les descripteurs. Formellement, on a :

$$\widehat{Y} = X(X^t D X)^{-1} X^t D Y \tag{4}$$

Une fois la matrice $\widehat{Y}$ calculée, la seconde étape consiste à effectuer et interpréter une simple Analyse en Composantes Principales (ACP) sur cette dernière. On projette ainsi dans le plan factoriel uniquement la part des abondances d'espèces expliquée par les descripteurs, puisque les axes factoriels sont des combinaisons linéaires des espèces, elles-même combinaisons linéaires des descripteurs. L'algorithme 1 résume cette procédure.

Algorithme 1 Analyse en composantes principales sur variables instrumentales

Entrée: Les matrices d'abondance Y et des descripteurs X

procédure ACPVI(X, Y)
    Définir une métrique D ▷ Dans la formulation classique, $D = \frac{1}{n} I_n$
    pour $i \in [\![1; p]\!]$ **faire**
        Calculer $\widehat{Y_i} = \beta_{i,0} + \sum_{k=1}^{p} \beta_{i,k} X_k$
    fin pour
    Former la matrice $\widehat{Y}$ à partir des colonnes $\widehat{Y_i}$
    ACP($\widehat{Y}$)
    Inclure les descripteurs comme variables supplémentaires
fin procédure

Sortie: Les axes factoriels donnés par les vecteurs propres de l'ACP

Cette méthode permet ensuite d'avoir accès à un certain nombre d'informations :

- Le pourcentage d'inertie expliqué par les descripteurs, en faisant le rapport entre la somme des valeurs propres de l'ACPVI et la somme des valeurs propres d'une simple ACP sur Y. Ceci nous donne une idée de l'interprétabilité des résultats de l'ACPVI : si cette dernière explique un très faible pourcentage de l'inertie totale, alors les descripteurs ne permettent simplement pas d'expliquer les abondances, et inversement.

- La mise en évidence des espèces à forte variance, donc ayant un indice d'abondance souvent élevé, en calculant la covariance entre les espèces et les axes factoriels.
- Le ou les descripteurs les plus importants pour expliquer les abondances des espèces, en calculant la corrélation entre chaque descripteur et les axes factoriels, ce qui revient à considérer les descripteurs comme des variables supplémentaires de l'ACP. En effet, les axes factoriels résument en quelque sorte la structure de $\widehat{Y}$. Les descripteurs ayant une corrélation forte avec les axes factoriels (en particulier les premiers) sont donc ceux ayant le plus de pouvoir explicatif sur la structure de $\widehat{Y}$. Si le pourcentage d'inertie expliqué est suffisamment grand, il est raisonnable de généraliser cela à l'échelle de Y, donc des abondances.
- Les espèces fortement liées à un facteur donné, en comparant la corrélation espèces/axes factoriels à celle descripteurs/axes factoriels : si, dans le plan formé par les deux premiers axes de l'ACPVI, une espèce est située dans la même région qu'un descripteur (coordonnées données par les corrélations), alors l'abondance de l'espèce en question est probablement liée à ce dernier. Encore une fois, cette interprétation est à mettre en perspective du pourcentage d'inertie expliqué par l'ACPVI.

5.2 Analyse par ACPVI successives

Si cette méthode dispose de nombreux avantages, il faut garder en tête qu'il est un problème que nous n'avons pas encore réglé : celui de la corrélation des différents descripteurs de X. Ce point est tout particulièrement important lorsque l'on interprète l'ACPVI afin de trouver les descripteurs les plus importants pour expliquer la matrice d'abondance. Une erreur facile à commettre, et pourtant essentielle à éviter, serait de prendre les k descripteurs les plus corrélés avec les axes, de les ordonner par force de la corrélation, et de les déclarer descripteurs les plus importants, dans cet ordre là. Or, nous n'avons à ce niveau là pas de moyen de savoir si la forte corrélation d'un descripteur avec les axes est "réelle", ou bien si elle est due à la corrélation de celui-ci avec un autre descripteur important. Prenons un exemple : à l'issue de l'ACPVI, on remarque que les deux facteurs les plus corrélés avec les axes sont le type de culture et l'altitude. Dans ce cas, est-ce que l'altitude influe directement sur l'abondance des espèces, ou bien influe-t-elle uniquement sur le type de culture, qui, lui, influe directement sur les espèces ? Dans le premier cas, l'altitude est un facteur de première importance, là où, dans le second, il ne constitue qu'une information redondante avec le type de culture.

Pour éviter cela, il faut veiller à n'interpréter qu'un seul descripteur majeur à chaque ACPVI. Pour établir une hiérarchie entre les descripteurs, il faut ensuite effectuer une seconde analyse dite orthogonale au descripteur trouvé lors de la première ACPVI. Il est donc nécessaire de calculer une seconde ACPVI sur la part d'information des autres facteurs indépendante du premier descripteur. Pour cela, on calcule les résidus de la régression des autres facteurs sur le premier, puis on réalise l'ACPVI sur ces résidus plutôt que sur les facteurs eux-mêmes. Dans notre exemple, si on pose que la culture est la facteur majeur, on calcule les résidus des régressions de chaque autre facteur sur le type de culture, puis on réalise l'ACPVI sur cette nouvelle matrice de descripteurs. Cela permet d'éliminer complètement l'effet du type de culture sur les autres facteurs, et donc sur les représentations des espèces dans l'ACPVI. Ainsi, les facteurs et espèces fortement dépendants du type de culture se retrouveront projetés proche de l'origine sur la seconde ACPVI, laissant la place aux facteurs réellement explicatifs. L'algorithme 2 résume cette procédure d'analyse par ACPVI successives (dites aussi ACPVI orthogonales).

Algorithme 2 Analyse par ACPVI successives

Entrée: Les matrices d'abondance Y et des descripteurs X

procédure ACPVI_ORTHO(X, Y)
    ACPVI(X, Y)
    Effet majeur du facteur type de culture
    pour $i \in [\![1; p']\!]$ **faire** ▷ p' vaut p moins le nombre de modalités du type de culture
        Calculer X'_i comme les résidus de la régression de X_i par les types de culture
    fin pour
    Former la matrice X' à partir des colonnes X'_i
    ACPVI(X', Y)
    Effet majeur du facteur désherbage
    pour $i \in [\![1; p'']\!]$ **faire** ▷ p'' vaut p moins le nombre de modalités des type de culture et de désherbage
        Calculer X''_i comme les résidus de la régression de X_i par les types de culture et de désherbage
    fin pour
    Former la matrice X'' à partir des colonnes X''_i
    ACPVI(X'', Y)
    Effet majeur du facteur pluviométrie
    Etc...
fin procédure

Sortie: Les descripteurs ayant le plus fort pouvoir explicatif

5.3 Application et interprétations

Une fois la méthode choisie et définie, nous l'avons mise en application sur des données issues de La Réunion, pour lesquelles nous disposions de suffisamment de descripteurs potentiellement explicatifs des abondances. Dans ce cadre là, nous disposons de 481 relevés floristiques, recensant 273 espèces adventices distinctes. Pour chaque relevé, nous disposions en plus de 6 descripteurs d'intérêt, dont deux quantitatifs (l'altitude de la parcelle et la pluviométrie moyenne) et 4 qualitatifs (le type de culture, la culture précédente, le type d'irrigation et le type de désherbage utilisé). La table 2 décrit la composition des différents descripteurs. Les analyses ont été réalisées via le package `amatrop`, développé au cours de ce stage (voir la conclusion et l'Annexe 4 pour plus de détails). Les Figures 5 et 6 contiennent les sorties, brutes et graphiques, d'une analyse par ACPVI.

Table 2: Noms et modalités des différents descripteurs utilisés pour les analyses multivariées

alt	**pluvio**	**cult**	**preced**	**irrig**	**desherb**
Quantitatif	Quantitatif	Ananas	Ananas		Mecanique
		Arbo.	Arbo.		Non_desherb.
		Canne	Canne		prelevee
		Jachere	Friche	Irrigue	postlevee
		Lentilles	Jachere	Non_irrig.	post_meca
		Maraich.	Lentilles		pre_meca
		Prairie	Maraich.		pre_post
			Prairie		pre_post_meca

L'interprétation de ces résultats permet de tirer bon nombre d'informations. Tout d'abord, on constate que le pourcentage d'inertie de la matrice d'abondance expliqué par l'ACPVI est de presque 44%, soit une valeur plutôt forte : il nous manque des éléments, mais l'interprétation de l'ACPVI a tout de même un intérêt certain. De plus, quand on décompose l'inertie expliquée par axe factoriel, on constate que l'axe 1 contient à

```
This is PCAIV without center and without scale

Inertia of the abundance matrix explained by given factors (%): 43.81

Inertia per axis (%):
Axis 1 Axis 2 Axis 3 Axis 4 Axis 5
 29.83   4.84   3.41   1.49   0.75

Cumulative inertia (%):
Axis 1:1 Axis 1:2 Axis 1:3 Axis 1:4 Axis 1:5
   29.83    34.67    38.08    39.57    40.32


10 most correlated descriptors with each factorial axes:
                        Axis  1                          Axis  2                              Axis  3

1              cult.Prairie  0.798 ***         preced.Canne  0.521 ***                  pluvio  0.772 ***
2                       alt  0.790 ***        cult.Maraich. -0.517 ***           irrig.Irrigue -0.691 ***
3            preced.Jachere  0.780 ***           cult.Canne  0.504 ***        irrig.Non_irrig.  0.691 ***
4              preced.Canne -0.734 ***      preced.Maraich. -0.504 ***    desherb.Non_desherb.  0.491 ***
5                cult.Canne -0.664 ***                  alt -0.490 ***            preced.Canne  0.407 ***
6     desherb.Non_desherb. -0.609 ***    desherb.Mecanique -0.354 ***           cult.Maraich. -0.401 ***
7        desherb.post_meca  0.582 ***        irrig.Irrigue  0.308 ***         preced.Maraich. -0.376 ***
8                   pluvio  0.251 ***     irrig.Non_irrig. -0.308 ***              cult.Canne  0.365 ***
9        desherb.postlevee  0.228 ***     desherb.prelevee  0.285 ***       desherb.Mecanique -0.353 ***
10          cult.Lentilles  0.224 ***               pluvio -0.285 ***                     alt -0.235 ***


10 most correlated species with each factorial axes:
         Axis  1             Axis  2             Axis  3

1     HOLLA  0.560 ***    STEME -0.516 ***    KYLEL 0.569 ***
2     PANMA -0.415 ***    GASPA -0.510 ***    RUBAC 0.456 ***
3     SIKOR -0.382 ***    PLALA -0.443 ***    SIKOR 0.436 ***
4     LISGU -0.358 ***    CASOC  0.442 ***    SETPF 0.357 ***
5     SOLAM -0.310 ***    PANMA  0.435 ***    OXACB 0.350 ***
6     MOMCH -0.296 ***    RAPRA -0.431 ***    COMDI 0.343 ***
7     CRIHA -0.293 ***    FUMMU -0.425 ***    UOUJA 0.339 ***
8     CYPRO -0.292 ***    DEMVI  0.418 ***    MIMPU 0.335 ***
9     PASPA -0.282 ***    MOMCH  0.408 ***    CRSCR 0.335 ***
10    LANCA -0.280 ***    CYNDA  0.403 ***    DRYCO 0.326 ***


Signif. codes for correlation test's p values:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5: Sorties brutes de la fonction `PCAIV()` du package `amatrop`

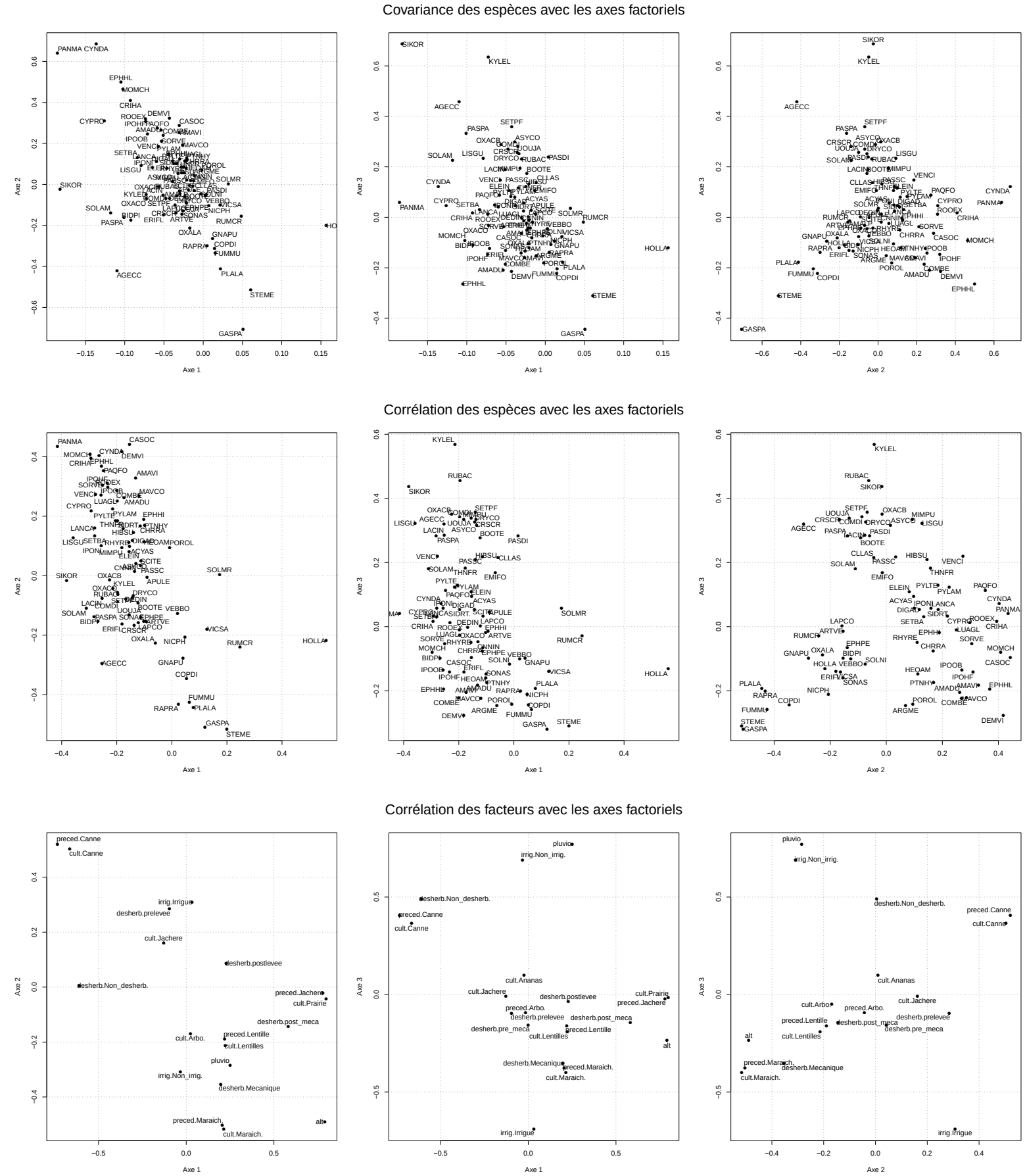

Figure 6: Sorties graphiques de la fonction `PCAIV()` du package `amatrop`

lui seul près de 30% de l'information. Il est donc plus intéressant de se pencher sur celui-ci, quitte à reléguer les autres axes factoriels à un rôle plus secondaire.

Il est ensuite nécessaire de regarder les facteurs les plus corrélés aux axes, puisque c'est ici que l'on trouvera les principaux facteurs explicatifs des abondances. Dans la corrélation à l'axe 1, on remarque surtout un effet fort de la culture (cult) (la flore de la canne à sucre est très différente de la flore du maraichage et des prairies) et de l'altitude (alt) (la flore de basse altitude est très différentes de la flore d'altitude), et dans une moindre mesure de la culture précédente (preced) et du désherbage (desherb). Ainsi, l'altitude, qui conditionne la culture (on trouve en effet de la canne à sucre à basse altitude et du maraichage et des prairies à haute altitude), apparait comme le facteur le plus explicatif de la présence et de l'abondance des espèces adventices de La Réunion.

On peut enfin étudier les sorties graphiques, et comparer les position des espèces et des facteurs sur les graphiques de corrélation espèces/axes et facteurs/axes : en effet, les espèces situées dans la même région qu'un facteur sont probablement fortement liées à celui-ci. on remarque ainsi que, sur l'axe 1, les espèces PANMA ou CYNDA sont situées dans la même zone que les facteurs `cult.Canne` et `preced.Canne`, et on en déduit que ces espèces sont fortement liées à la culture de la canne à sucre. De même, les espèces GASPA et STEME semblent être liées aux cultures maraîchères. HOLLA, elle, semble être liée aux cultures en jachère ou de prairie, et semble dépendante de l'altitude ; KYLEL, RUBAC et SIKOR semblent dépendantes de la pluviométrie, et plutôt liée aux cultures non irriguées dans les régions à forte pluviométrie.

Pour mettre en évidence les autres facteurs explicatifs de la matrice d'abondance, il convient ensuite de travailler selon les principes de l'analyse par ACPVI successives. Ces nouvelles analyses se font via la fonction `PCAIV_ortho()` du package `amatrop`. À l'issue de celles-ci, on constate que, contrairement à ce qu'on pouvait penser au regard des premiers résultats, l'importance de la culture diminue fortement lorsque l'on effectue l'analyse orthogonalement à l'altitude. Cela indique que l'effet du type de culture dépendait en fait largement de l'altitude, et que c'est bien ce dernier qui influait sur l'abondance des espèces. Au contraire, il ressort de cette seconde analyse que les différentes pratiques de désherbage sont essentielles. Une troisième analyse nous apprend que la pluviométrie joue également un rôle important. Puis viennent les types de culture et d'irrigation, alors que la culture précédente semble jouer un rôle extrêmement limité. .

5.4 Profils écologiques

Il est une autre métode, couramment utilisée en écologie, permettant d'analyser le lien entre une espèce et les différentes modalités d'un descripteur. Cette métode, nommée profils écologiques [9], repose sur le calcul de la fréquence d'une espèce en fonction des modalités d'un descripteur de milieu, et requiert donc de travailler en présence/absence. De manière générale, le calcul des profils écologiques se fait pour une espèce E et un descripteur L donnés, et fait référence à l'analyse de plusieurs quantités, issues pour beaucoup de la théorie de l'information [2].

Les profils d'ensemble contiennent simplement le nombre de relevés disponibles par modalité du descripteur considéré. Les profils d'ensemble permettent de se faire une idée de la qualité d'échantillonnage, à travers les différentes modalités. Du point de vue des notations, on note $R(K)$ la valeur du profil d'ensemble pour la modalité K, et $N_R = \sum_K R(K)$ le nombre total de relevés.

L'entropie-facteur permet par ailleurs de quantifier la qualité de l'échantillonnage. En effet, si un descripteur est bien échantillonné, alors les différentes modalités de ce dernier sont équiprobables, et l'indétermination, ou la quantité d'information, est grande. Cette indétermination est estimée par l'entropie-facteur $\widehat{H}(L)$:

$$\widehat{H}(L) = \sum_K \frac{R(K)}{N_R} \log_2 \left(\frac{N_R}{R(K)} \right) \tag{5}$$

De plus, on sait que l'entropie-facteur atteint sa valeur maximal quand l'échantillonnage comporte un nombre égal de relevés dans chaque modalités du descripteur. Dans ce cas là $\frac{R(K)}{N_R} = \frac{1}{N_K}$ (avec N_K le nombre de modalités), d'où $R(K) = \frac{N_R}{N_K}$, et finalement :

$$H_{max}(L) = \log_2(N_K) \tag{6}$$

On peut dès lors quantifier la qualité d'échantillonnage, pour un descripteur L donné, comme :

$$Q(L) = \frac{\widehat{H}(L)}{H_{max}(L)}, \tag{7}$$

qui donne des valeurs proches de 1 pour un descripteur bien échantillonné.

Génrélament, on calcule ensuite les profils écologiques proprement dits. Il s'agit de calculer la fréquence corrigée de chaque espèce en fonction des modalités du facteur considéré. En effet, une fréquence absolue est insuffisante, car elle est trop dépendante de l'échantillonnage. La fréquence relative (à chaque modalité) permet de pallier ce problème, mais donne trop de poids aux espèces globalement fréquentes, et trop peu aux espèces rares. C'est pourquoi on utilise une fréquence dite corrigée, en divisant les fréquences relatives de l'espèce pour chaque modalité par la fréquence relative globale de l'espèce, calculée dans l'ensemble des relevés. Il est ensuite de coutume de multiplier le résultat par 100, et ainsi, une fréquence corrigée de 100 (ou proche de 100) dans une des modalités indique une présence moyenne dans celle-ci (on la qualifie parfois de due au hasard, dans le sens non-liée à la modalité), une fréquence corrigée largement plus grande que 100 suggère que l'espèce est très présente dans cette modalité (donc liée à la modalité), et inversement. Formellement, on a :

$$C(K) = \frac{U(K)/R(K)}{U(E)/N_R}, \tag{8}$$

avec :

- $C(K)$: la fréquence corrigée
- $U(K)$: la fréquence absolue de l'espèce dans la modalité K du descripteur
- $U(E) = \sum_K U(K)$: la fréquence absolue de l'espèce dans tous les relevés

À partir des profils écologiques corrigés, on peut calculer une quantité nommée entropie-espèce, quantifiant cette fois-ci la qualité de l'échantillonnage de l'espèce, valant 1 au maximum si l'espèce est présente dans la moitié des relevés, et défini comme :

$$\widehat{H}(E) = \frac{U(E)}{N_R} \log_2\left(\frac{N_R}{U(E)}\right) + \frac{N_R - U(E)}{N_R} \log_2\left(\frac{N_R}{N_R - U(E)}\right), \tag{9}$$

La dernière quantité classiquement calculée lors d'une analyse par profils écologiques est l'information mutuelle [14] entre l'espèce et le descripteur. Cette dernière estime la quantité d'information apportée par une espèce relativement à un descripteur. Elle permet donc, pour une espèce donnée, de distinguer les descripteurs les plus importants pour la présence de l'espèce. L'information mutuelle vaut :

$$\widehat{I}(L,\, E) = \sum_K \frac{U(K)}{N_R} \log_2\left(\frac{U(K)}{R(K)} \cdot \frac{N_R}{U(E)}\right) + \sum_K \frac{V(K)}{N_R} \log_2\left(\frac{V(K)}{R(K)} \cdot \frac{N_R}{V(E)}\right), \tag{10}$$

avec $V(K) = R(K) - U(K)$ le nombre de fois où l'espèce a été absente dans un relevé effectué dans la modalité K du descripteur. Plus l'information mutuelle est forte, plus le facteur est important pour la présence ou l'absence de l'espèce. Un exemple complet de profils écologiques pour les 20 espèces présentant la plus forte information mutuelle avec le descripteur "type de culture" est presenté en Annexe 2. Tout comme pour l'ACPVI, les analyses ont été réalisée via le package `amatrop`.

5.5 Apprentissage de la présence/absence des espèces

Le fait d'avoir à disposition des données ayant plusieurs facteurs environnementaux bien renseignés, et surtout la possibilité d'en avoir de plus en plus à mesure que les jeux de données actuels seront complétés par de nouveaux, nous a poussé à vouloir mettre en place un cadre d'apprentissage machine de la présence (ou l'absence) des espèces adventices. L'idée est donc d'entraîner des modèles d'apprentissage sur les données disponibles, afin d'être par la suite en mesure de prédire, pour une nouvelle parcelle que l'on souhaiterait cultiver, une "liste" d'espèces potentiellement présentes, ainsi que, pour chacune de ces espèces, une probabilité estimée de présence.

Pour cela, le cadre d'apprentissage suivant a été mis en place :

- Un modèle distinct par espèce a été entraîné (les espèces très rares, et non directement liées à un facteur rare sont enlevées).
- Pour chaque espèce, un modèle de forêt aléatoire classique [6], opérant en classification binaire, a été entraîné à prédire la présence ou l'absence de celle-ci, sans tenir compte de l'abondance.
- Pour chaque espèce, un second modèle, de forêt de probabilité cette fois [26], à été entraîné. Ce second modèle présente l'intérêt d'estimer une probabilité de présence, interprétable comme telle.
- Pour chaque espèce, les hyper-paramètres des modèles ont été choisis par *grid search* et validation croisée (5 *folds*). Au cours de cette sélection de modèle, les meilleurs modèles ont été sélectionnés selon le critère de précision (*accuracy score*) pour les forêts de classification et d'aire sous la courbe ROC pour les forêts de probabilité.
- Une fois les meilleurs modèles trouvés pour chaque espèces, ceux-ci ont été entraînés sur toutes les données de La Réunion présentées auparavant. Enfin, il nous fut possible d'extraire, toujours par espèce, deux mesures de la qualité des prédictions : le score de précision pour les modèles de classification, et le score de Brier [7] pour les forêts probabilistes. Ce dernier correspond à une mesure de l'adéquation entre la probabilité prédite et la classe observée, et vaut dans sa formulation classique :

$$BS = \frac{1}{n}\sum_{t=1}^{n}(f_t - o_t)^2, \tag{11}$$

où f_t est la probabilité prédite, o_t est l'observation binaire, et N le nombre de prédictions effectuées. Notons que, comme nous avons utilisé l'ensemble des données de La Réunion pour entraîner ces modèles, les scores de précision et de Brier sont estimés par leurs valeurs dites *Out Of Bag*, estimation permise par l'utilisation du *bootstrap* dans les modèles de forêts aléatoires. L'ensemble des résultats pour les 85 espèces les plus présentes est présenté en Annexe 3.

6 Conclusion

Au fil de ce projet, nous avons ainsi effectué un travail complet sur un volume de données conséquent concernant la présence et l'abondance d'espèces de plantes adventices au sein d'une diversité de systèmes de cultures différents, tous situés dans des climats tropicaux. Les données sont à la fois nombreuses et disparates, détaillant les situations de nombreux pays, surtout africains, mais aussi d'Asie ou d'Amérique du Sud, et sont étalées sur plus de 30 ans. Les principaux intérêts de ce projet sont :

- D'une part le rassemblement et la mise à disposition d'une large masse de données concernant un domaine encore trop largement méconnu.
- D'autre part la production d'une chaine complète d'analyse du comportement des adventices en fonction de leur abondance et des facteurs de milieu.

Le projet Amatrop est ainsi le premier rassemblement publique de données permettant d'analyser la répartition et l'abondance des espèces adventices dans les cultures tropicales à une échelle globale. Les formats standards choisis pour les données permettent par ailleurs une grande évolutivité, facilitant ainsi l'ajout continu de nouveaux jeux de données et un travail collaboratif constructif entre malherbologues tropicalistes. Ainsi, à l'issue de ce travail, deux publications scientifiques ont été envisagées. Une première concerne la collecte, l'homogénéisation, et la mise à disposition des données, et consiste en un *datapaper* soumis à l'*Open Data Journal for Agricultural Research*[5] (ODJAR), et est en cours de traitement. Un second article concernant l'interprétation des résultats des différentes analyses est également prévu. En outre, le travail effectué au cours de ce stage fera l'objet de plusieurs présentations orales au sein de différentes unités du CIRAD et de ses patenaires.

Le travail réalisé présente enfin l'avantage de couvrir la majorité des tâches relatives à tout projet d'analyse de données : rassemblement et récupération des données depuis des tableurs ou des bases de données anciennes, gestion et nettoyage des données, standardisation et homogénéisation de tous les jeux de données sur une base commune pour pouvoir conduire des analyses conjointes regroupant plusieurs jeux de données, stockage et mise à disposition pour utilisation publique des données, mise en œuvre d'analyses statistiques plus ou moins complexes, et écriture d'un cadre d'apprentissage machine de la présence des espèces.

Chacunes de ces tâches à d'abord été analysées en partenariat avec les malherbologues, puis a dû être implémentée. Or, pour effectivement conduire ces analyses, plusieurs contraintes importantes devaient être considérées. Tout d'abord, les analyses devaient être conduites sous `R` [38] afin de permettre à un maximum d'utilisateurs de pouvoir travailler avec, tout en ayant un accès aisé au code source. Ensuite, le tout devait être rassemblé sous la forme de scripts automatisés au maximum et faciles d'utilisation, permettant aisément d'effectuer les analyses souhaitées pour quelqu'un n'étant pas familier avec `R`. Enfin, les scripts en question se devaient d'être évolutifs, l'objectif étant d'ajouter des jeux de données au cours du temps.

À la vue de ces contraintes, et après avoir considéré plusieurs options, il a été décidé de développer un package `R` rassemblant des fonctions permettant de :

- traiter les données
- générer des fichiers propices au stockage publique et à l'analyse à partir des sources de données en Excel
- concaténer les données disponibles selon des filtres définis par l'utilisateur afin de permettre à ce dernier de conduire des analyses selon un niveau de finesse contrôlé
- mettre en œuvre les analyses souhaitées.

Le package, nommé `amatrop` à l'image du projet, a été développé sous la version 3.6.1 de `R` et est compatible avec la version 2.10 ou toute version plus récente, jusqu'à la 4.0. Le package a été développé à l'aide des outils `devtools` et `roxygen2`. Pour l'instant, le package n'a pas vocation à être rendu public via le CRAN, puisque le projet Amatrop reste principalement interne au CIRAD. Son installation se fait donc uniquement grâce au fichier source. Le package comprend de plus un certain nombre de dépendances devant être installées a priori.

Un total de 14 fonctions ont été exportées via `amatrop`, avec un soin particulier apporté à l'évolutivité

[5] https://odjar.org/

et l'automatisation de celles-ci, ainsi qu'à leurs performances : deux concernent la gestion des données ; une permet de combiner et filtrer l'ensemble des données selon un niveau de finesse dont le choix est laissé à l'utilisateur, en vue d'analyses futures ; cinq sont utilisées pour générer les analyses simples relatives à l'abondance et à la fréquence des espèces ; une est relative aux profils écologiques ; deux permettent d'efféctuer les analyses reposant sur l'ACPVI ; et quatre servent enfin a entraîner et utiliser à des fins de prédictions les modèles d'apprentissage machine. Deux documentations ont de plus été rédigées à propos du package, dont une en français, très détaillée, à l'intention des utilisateurs du CIRAD. Une seconde documentation, plus technique, en anglais et générée grâce aux outils de création de package propres à `R`, a également été rédigée, et est présentée en Annexe 4.

La création d'une chaîne complète d'analyses automatisées permet enfin d'envisager un travail de modélisation plus profond. En effetdes projets de modélisation de la présence et de l'abondance des mauvaises herbes sur l'île de La Réunion sont envisagés par l'équipe du Cirad. Cette chaîne de traitement et d'analyse peut en ce sens servir de point de départ à ce travail de modélisation, puisqu'elle permet de mettre en évidence à la fois les espèces et les facteurs importants, ainsi que leurs interactions. C'est pourquoi le travail effectué durant ce stage n'est qu'une porte d'entrée, un point de départ pour d'autres projets. Malgré la fin de ce stage, le projet Amatrop a vocation à évoluer encore beaucoup, et n'est pas un "produit fini" : de nouveaux jeux données sont vouées à être ajoutés (certains ont été ajoutés durant l'écriture même de ce rapport), concernant d'autres régions tropicales du monde, permettant de donner petit à petit aux analyses réalisées une échelle de plus en plus globale.

Annexes

Annexe 1 : Informations supplémentaires sur les 25 jeux de données qui composent l'étude

Annexe 1.1 : Liste des abréviation utilisées au fil du projet

Table 3: Liste des abréviations utilisées pour les auteurs, pays, culture de score floristique

	Nom complet	Abréviation
Pays	Bénin	BEN
	Cameroun	CAM
	Côte d'Ivoire	CDI
	Guinée	GUI
	Guyane	GUY
	Madagascar	MAD
	Île Maurice	MAU
	Île de La Réunion	RUN
	Viêt Nam	VIE
Auteur	Baillif Stéphane	BAI
	Boraud Maxime	CAM
	Gnahoua G.M.	GNA
	Ipou Ipou Joseph	IPO
	Kouame Sabine	SAB
	Le Bourgeois Thomas	LEB
	Marnotte Pascal	MAR, MA2, MA3
	Rafenomandjato Anstsa	ANT
	Randriamampianina J.A.	JAR, JA2
	Stevoux Véronique	STE
	Tehia Kouakou Etienne	TEH
	Toure Awa	AWA
	Vall Eric	VAL
	Viaud Pauline	VIA
Culture	Coton	COT
	Diverses	DIV
	Jachère	JAC
	Vivrier	VIV
	Riz	RIZ
	Canne à sucre	CAN
	Maraîchage	MAR
Notation floristique	0-1 (Présence/absence)	PA
	1-9 CEB (Abondance)	AB

Annexe 1.2 : Table de convertion des scores d'abondance

Table 4: Conversion des différents scores d'abondance

Dominance	**1-5 Braun Blanquet**	**Pourcentage de recouvrement**	**Observation**	**1-9 CEB**
	1	1-3	Espèce présente mais rare	1
3	1	4-10	Moins d'un individu par m^2	2
2	2	11-20	Au moins un individu par m^2	3
1	3	21-40	30% de recouvrement	4
	3	41-60	50% de recouvrement	5
	4	61-75	70% de recouvrement	6
	4	76-90	Très fort taux de recouvrement	7
	5	91-99	Très peu de sol apparent	8
	5	100	Recouvrement total	9

Annexe 1.3 : Situation géographique et temporelle des jeux de données

Table 5: Présentation géographique et temporelle des jeux de données

Etude	**Pays**	**Mono-site/ Multi-sites**	**Auteur**	**Année**
BEN-MAR-2013-DIV-AD	Bénin	Mono-site	Marnotte Pascal	2013
CAM-LEB-1988-DIV-AD	Cameroun	Multi-site	Le Bourgeois Thomas	1988-1991
CAM-MAR-1999-MSK-AD	Cameroun	Multi-sites	Marnotte Pascal	1999
CAM-VAL-2001-COT-AD	Cameroun	Mono-site	Vall Eric	2001
CDI-AWA-2011-VIV-AD	Côte d'Ivoire	Mono-site	Touré Awa	2011-2014
CDI-BOR-1997-CAN-PA	Côte d'Ivoire	Multi-sites	Boraud Maxime	1997
CDI-GNA-1997-JAC-AD	Côte d'Ivoire	Mono-site	Gnahoua Guy Modeste	1997
CDI-IPO-2009-DIV-PA	Côte d'Ivoire	Multi-sites	Ipou Ipou Joseph	2009-2018
CDI-MAR-1991-CAN-AD	Côte d'Ivoire	Mono-site	Marnotte Pascal	1991-1992
CDI-SAB-2015-VIV-AD	Côte d'Ivoire	Mono-site	Kouamé Sabine	2015
CDI-TEH-2014-COT-AD	Côte d'Ivoire	Multi-sites	Téhia Kouakou Etienne	2014
GUI-MAR-1996-COT-AD	Guinée	Mono-site	Marnotte Pascal	1996-1997
GUY-LEB-2018-DIV-AD	Guyane	Multi-sites	Le Bourgeois Thomas	2018
MAD-ANT-2017-RIZ-AD	Madagascar	Multi-sites	Rafenomanjato Antsa	2017
MAD-JAR-1998-DIV-AD	Madagascar	Multi-sites	Randriamampianina J.A.	1998
MAD-JA2-1999-DIV-AD	Madagascar	Multi-sites	Randriamampianina J.A.	1999
MAD-RAK-2015-RIZ-AD	Madagascar	Multi-sites	Rakotonirina Fetisoa	2015
MAU-MAR-2019-MAR-AD	Île Maurice	Multi-sites	Marnotte Pascal	2019
RUN-BAI-2017-CAN-AD	La Réunion	Mono-site	Baillif Stephane	2017-2018
RUN-LEB-2003-DIV-AD	La Réunion	Multi-sites	Le Bourgeois Thomas	2003-2006
RUN-MAR-2016-CAN-AD	La Réunion	Multi-sites	Marnotte Pascal	2005-2016
RUN-MA2-2019-CAN-AD	La Réunion	Mono-site	Marnotte Pascal	2019-2020
RUN-MA3-2019-CAN-AD	La Réunion	Mono-site	Marnotte Pascal	2019-2020
RUN-VIA-2019-CAN-AD	La Réunion	Mono-site	Viaud Pauline	2019-2020
VIE-STE-1999-RIZ-AD	Viet Nam	Mono-site	Stevoux Véronique	1999

Annexe 1.4 : Outils de citation des études

Table 6: Liste des liens et outils de citation par études

Etude	DOI	Citation
BEN-MAR-2013-DIV-AD	https://doi:10.18167/DVN1/X98KVY	[32]
CAM-LEB-1988-DIV-AD	https://doi:10.18167/DVN1/1HCKBU	[20]
CAM-MAR-1999-MSK-AD	https://doi:10.18167/DVN1/H7AJUK	[34]
CAM-VAL-2001-COT-AD	https://doi:10.18167/DVN1/9PDXYC	[15]
CDI-AWA-2011-VIV-AD	https://doi:10.18167/DVN1/ZH4W5M	[48]
CDI-BOR-1997-CAN-PA	https://doi:10.18167/DVN1/Y5SRMR	[4]
CDI-GNA-1997-JAC-AD	https://doi:10.18167/DVN1/RRQWXJ	[13]
CDI-IPO-2009-DIV-PA	https://doi:10.18167/DVN1/V0QGKT	[16]
CDI-MAR-1991-CAN-AD	https://doi:10.18167/DVN1/3KAVBK	[30]
CDI-SAB-2015-VIV-AD	https://doi:10.18167/DVN1/VCLIB5	[17]
CDI-TEH-2014-COT-AD	https://doi:10.18167/DVN1/TCCSRC	[47]
GUI-MAR-1996-COT-AD	https://doi:10.18167/DVN1/EVCQFR	[29]
GUY-LEB-2018-DIV-AD	https://doi:10.18167/DVN1/TMZSMW	[18]
MAD-ANT-2017-RIZ-AD	https://doi:10.18167/DVN1/VSY7Y5	[39]
MAD-JAR-1998-DIV-AD	https://doi:10.18167/DVN1/O1EHV2	[40]
MAD-JA2-1999-DIV-AD	https://doi:10.18167/DVN1/O1EHV2	[41]
MAD-RAK-2015-RIZ-AD	https://doi:10.18167/DVN1/NQYPE	[42]
MAU-MAR-2019-MAR-AD	https://doi:10.18167/DVN1/VBE4VT	[31]
RUN-BAI-2017-CAN-AD	https://doi:10.18167/DVN1/GDNUCH	[3]
RUN-LEB-2003-DIV-AD	https://doi:10.18167/DVN1/UAHMEJ	[21]
RUN-MAR-2016-CAN-AD	https://doi:10.18167/DVN1/MDP0H5	[27]
RUN-MA2-2019-CAN-AD	https://doi:10.18167/DVN1/FWFAUY	[33]
RUN-MA3-2019-CAN-AD	https://doi:10.18167/DVN1/1HCKBU	[28]
RUN-VIA-2019-CAN-AD	https://doi:10.18167/DVN1/YWNUCG	[49]
VIE-STE-1999-RIZ-AD	https://doi:10.18167/DVN1/OFZEH1	[46]

Annexe 1.5 : Nombre de relevés par pays et par culture

Table 7: Nombre de relevés par pays et par culture

	arachide	jachère	riz	tubercule	vivrier	cotonnier	canne	arboriculture	maraîchage	Ananas	prairie	Total
Bénin	5	3	6	2	6	0	0	0	0	0	0	22
Cameroun	0	0	0	0	52	1395	0	0	0	0	0	1447
Côte d'Ivoire	0	64	0	135	305	619	265	0	0	0	0	1388
Guinée	0	0	0	0	0	110	0	0	0	0	0	110
Guyane	0	0	0	0	0	0	0	22	39	0	0	61
Madagascar	54	0	1319	118	282	226	0	0	0	0	0	1999
Maurice	0	0	0	0	0	0	0	0	82	0	0	82
Réunion	0	7	0	0	17	0	1008	5	84	25	38	1184
Vietnam	0	0	220	0	143	0	0	0	0	0	0	363
Total	59	74	1545	255	805	2350	1273	27	205	25	38	6656

Annexe 2 : Profils écologiques des 20 espèces ayant la plus forte information mutuelle avec le type de culture

Ecological profile of species with type = 'value':

Variable name: cult
Number of classes: 7
Number of readings: 481
Variable entropy: 1.695; Maximum entropy: 2.807; Quality of sampling: 0.604
Mean species entropy: 0.345

Classes and overall profiles:

Ananas	Arbo.	Canne	Jachere	Lentilles	Maraich.	Prairie
25	5	305	7	17	84	38

Corrected ecological profiles:

				..... CORRECTED PROFILES						
Species	Abs.frequency	Species.entropy	Mutual.info	Ananas	Arbo.	Canne	Jachere	Lentilles	Maraich.	Prairie
IUNEF	32	0.353	0.303	0	0	0	0	0	0	1266
AOXOD	31	0.345	0.290	0	0	0	0	0	0	1266
PANMA	265	0.993	0.281	123	73	133	182	11	30	0
DACGL	29	0.329	0.266	0	0	0	0	0	0	1266
GASPA	118	0.804	0.236	82	326	51	0	360	272	0
ERAMC.	26	0.303	0.232	0	0	0	0	0	0	1266
HRYRA	45	0.448	0.226	43	214	14	0	63	89	872
STEME	68	0.588	0.195	28	141	23	0	374	371	56
HOLLA	77	0.635	0.194	50	250	45	0	37	119	559
AGSTE	22	0.268	0.190	0	0	0	0	0	0	1266
PESCL	47	0.462	0.186	0	0	57	0	0	12	781
LISGU	192	0.970	0.172	130	0	136	72	15	30	0
MOMCH	186	0.963	0.169	124	52	133	222	0	31	0
ERIKA	25	0.295	0.149	0	0	6	0	226	46	1013
COPDI	116	0.797	0.145	66	166	68	0	171	262	0
SIGDF	17	0.221	0.142	0	0	0	0	0	0	1266
SOLAM	320	0.920	0.136	84	150	107	150	97	115	4
CYNDA	232	0.999	0.134	91	0	128	148	0	44	49
ACAMR	16	0.211	0.133	0	0	0	0	0	0	1266
BIDPI	329	0.900	0.132	123	88	107	104	146	101	8

Figure 7: Profils écologiques pour le facteur culture

Annexe 3 : Estimation de la qualité des prédictions de présence/absence des espèces

Table 8: Scores de précision et de Brier *Out of bag*

Espèce	**Accuracy**	**Brier score**	**Espèce**	**Accuracy**	**Brier score**
ACYAS	0.8835759	0.09841443	LACIN	0.7484407	0.16102495
AGECC	0.7671518	0.16265874	LANCA	0.5987526	0.22043719
AMADU	0.7151767	0.17754345	LAPCO	0.8794179	0.09394339
AMALP	0.7255717	0.19001395	LISGU	0.7318087	0.17353703
AMAVI	0.7962578	0.14658177	LUAGL	0.8648649	0.09882212
APULE	0.8461538	0.11954355	MAVCO	0.7463617	0.17167267
ARGME	0.8607069	0.09529264	MIMPU	0.8419958	0.10795869
ARTVE	0.8794179	0.09769079	MOMCH	0.7650728	0.16226586
ASYCO	0.8794179	0.08743612	NICPH	0.8898129	0.08346506
BIDPI	0.7713098	0.16647960	OXACB	0.7837838	0.16596250
BOOTE	0.8607069	0.10865579	OXACO	0.7546778	0.17608846
CASOC	0.8212058	0.12422397	OXALA	0.7255717	0.17833151
CHRRA	0.7629938	0.16692824	PANMA	0.7754678	0.15471642
CLLAS	0.8898129	0.08475710	PAQFO	0.8316008	0.10576377
CNNIN	0.8898129	0.09377460	PASDI	0.7505198	0.16980372
COMBE	0.7525988	0.17649564	PASPA	0.7900208	0.14635362
COMDI	0.7713098	0.16045947	PASSC	0.8856549	0.09369614
COPDI	0.8232848	0.13806543	PLALA	0.8316008	0.12186483
CRIHA	0.7318087	0.15988601	POROL	0.8108108	0.13636933
CRSCR	0.7837838	0.14602461	PTNHY	0.8690229	0.09896317
CYNDA	0.7006237	0.19267634	PYLAM	0.7713098	0.16397282
CYPRO	0.5779626	0.22824616	PYLTE	0.8066528	0.14583327
DEDIN	0.8607069	0.10925268	RAPRA	0.9376299	0.05119173
DEMVI	0.8357588	0.12392725	RHYRE	0.8794179	0.10023270
DIGAD	0.6798337	0.21110778	ROOEX	0.7775468	0.14838285
DRYCO	0.9189189	0.07100533	RUBAC	0.8648649	0.08814956
ELEIN	0.6486486	0.21213536	RUMCR	0.8711019	0.10725602
EMIFO	0.8565489	0.10906018	SCITE	0.8711019	0.10488879
EPHHI	0.7837838	0.15010560	SETBA	0.7255717	0.16613281
EPHHL	0.7754678	0.15483117	SETPF	0.8253638	0.14336173
EPHPE	0.8939709	0.08138767	SIDRT.	0.8731809	0.10531383
ERIFL	0.6507277	0.21188940	SIKOR	0.7130977	0.18849318
FUMMU	0.8274428	0.11616608	SOLAM	0.7297297	0.17706465
GASPA	0.8482328	0.10118144	SOLMR	0.7858628	0.15944646
GNAPU	0.8004158	0.15066860	SOLNI	0.8648649	0.11009553
HEOAM	0.9064449	0.07014381	SONAS	0.6465696	0.22143049
HIBSU	0.8939709	0.07924095	SORVE	0.7796258	0.15735045
HOLLA	0.9230769	0.06760192	STEME	0.9230769	0.06718814
IPOHF	0.7941788	0.13690207	THNFR	0.8835759	0.09196111
IPONI	0.7858628	0.15368081	UOUJA	0.8066528	0.13014759
IPOOB	0.7089397	0.18425142	VEBBO	0.8711019	0.10432975
KYLEL	0.8960499	0.07899463	VENCI	0.7422037	0.16581044
			VICSA	0.9209979	0.07609308

Annexe 4 : Documentation technique complète du package `amatrop`

Package 'amatrop'

August 17, 2020

Title Analize and process data from the Amatrop project

Version 0.0.0.9000

Description The package contains tools to ease the processing and the analysis of data from the Amatrop project. Their are tools for different kinds of analysis, according to user-defined filters.

License GPL-3

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.0

Imports openxlsx, corrplot, ggrepel, ggplot2, ggpubr, methods, ade4, caret, maptools, ranger, dplyr, e1071

Depends R (>= 2.10)

NeedsCompilation no

Author Benjamin Fayolle [aut, cre]

Maintainer Benjamin Fayolle <fayolle.benjamin@gmail.com>

R topics documented:

ADmoy *Compute mean abundance of species*

Description

This function can be used to compute the mean abundance value of weed species accross all flora readings, for given list of datasets.

Usage

```
ADmoy(data)
```

Arguments

data	a list of data.frames of matrix. Each data.frame or matrix is one dataset, filtered or not.

Value

a list of named numeric vectors, each corresponding to one of the entrie datasets.

cover_percentage *Convert flora abundance from 1-9 CEB scale to cover percentage*

Description

This function can be used to convert flora abundance from 1-9 CEB scale to cover percentage. The results are saved in .txt format according to 'output_path' parameter. The correspondance between 1-9 CEB scale and over percentage can be found in the metadata of most datasets.

Usage

```
cover_percentage(input_files, sep = " ", output_path = "auto")
```

Arguments

input_files	a list of strings. Each string is the absolute or relative path to one flora array, in .txt format. Must be a list.
sep	the field separator character. Values on each line of the file are separated by this character.
output_path	a character string vector of the same size than 'input_files' giving the path to the folders to store each text file created, or "auto". If "auto", the function will try to detect the path automatically from the input path. Default to "auto".

Value

The function does not return anything.

Examples

```
## Not run:
input_files <- list("../Data/DonneesFinales/BEN-MAR-2013-DIV-AD/BEN-MAR-2013-DIV-AD.xlsx",
                    "../Data/DonneesFinales/CAM-LEB-1990-DIV-AD/CAM-LEB-1990-DIV-AD.xlsx",
                    "../Data/DonneesFinales/CAM-MAR-1999-MSK-AD/CAM-MAR-1999-MSK-AD.xlsx",
                    "../Data/DonneesFinales/CAM-VAL-2001-COT-AD/CAM-VAL-2001-COT-AD.xlsx",
                     "../Data/DonneesFinales/CDI-AWA-2011-VIV-AD/CDI-AWA-2011-VIV-AD.xlsx"
)

cover_percentage(input_files)

## End(Not run)
```

`eco_profiles` *Compute ecological profiles*

Description

This function can be used to compute the ecological profiles of the species given in input, with regards to one categorical factor. This function is usefull in order to get a sense of the relationship between species and the different modalities of a factor.

Usage

```
eco_profiles(
  variable,
  species,
  factors,
  file = NULL,
  type = "test",
  mode = "presence",
  sort = FALSE,
  save_csv = FALSE
)
```

Arguments

`variable`	a character string matching one of the factor's name, i.e. matching a value in `colnames(factors)`, giving the factor by which the profiles will be calculated.
`species`	a matrix or data.frame containing the presence/absence values per species. Presence/absence should be described by 1 and 0 respectively. Each column should correspond to a single species.
`factors`	a matrix or data.frame containing the descriptors. Each column should correspond to a descriptor. Descriptors must be categorical.
`file`	a character string giving a valid path to the location the results should be stored, in a `.txt` format. If `NULL`, the results are not saved on the computer, but displayed instead on the console.

type	a character string, one of `c("test","value")`. If `"test"`, the corrected profiles will be replaced by a code reflecting the significance of the relationship between the species and the modalities. If "value", the corrected profiles are displayed.
mode	one of `c("presence","cover")`, indicating wether the profiles should be in presence/absence or in cover percentage. If the latter, then type can only be `"value"`.
sort	logical indicating if the results should be sorted according to the mutual information criteria.
save_csv	logical indicating if the ecological profiles should be saved in a `.csv` format. If `TRUE`, file must be filled, and the csv will be saved at the same location.

Value

Does not return anything. Instead, prints the results to a txt file or the console.

Examples

```
## Not run:
fac <- read.csv("D:/Documents/Amatrop/var.csv", row.names = 1, stringsAsFactors = TRUE)
ab <- read.csv("D:/Documents/Amatrop/ab.csv", row.names = 1)


PA <- as.data.frame(1 * (ab != 0)) ## Convert abundance to presence/absence

## Convert numerical factors to categorical
fac_quali <- fac
fac_quali$alt[fac$alt <= 400] <- "alt1"
fac_quali$alt[fac$alt > 400 & fac$alt <= 1000] <- "alt2"
fac_quali$alt[fac$alt > 1000 & fac$alt <= 1500] <- "alt3"
fac_quali$alt[fac$alt > 1500] <- "alt4"
fac_quali$pluvio[fac$pluvio <= 1000]  <- "pluvio1"
fac_quali$pluvio[fac$pluvio > 1000 & fac$pluvio <= 2000]  <- "pluvio2"
fac_quali$pluvio[fac$pluvio > 2000 & fac$pluvio <= 3500]  <- "pluvio3"
fac_quali$pluvio[fac$pluvio > 3500]  <- "pluvio4"

eco_profiles("cult", PA, fac_quali, file = "./Amatrop/test.txt", type = "test", sort = T, save_csv = F)

## End(Not run)
```

`filter_datasets` *Filter original data per factors*

Description

This function can be used to select desired factors and use them to filter the original datasets. Original datasets should be either Excel files (with `.xlsx` extension) or text files (see Arguments below). The function asks for desired factors to be used as filters, and if a filter is selected, it asks for which values to keep. See Details for more information on how to answer correctly to the function's questions. If not a single filter is selected, returns a dataset per study.

Usage

```
filter_datasets(
  input_txt_factors = NULL,
  input_txt_flora = NULL,
  input_xlsx = NULL,
  input_type = "txt",
  concatenate_filters = FALSE
)
```

Arguments

input_txt_factors	a list of strings, or NULL. Each string is the absolute or relative path to one dataset's factors, in text format. Must be a list.
input_txt_flora	a list of strings, or NULL. Each string is the absolute or relative path to one dataset's flora, in text format. Must be a list.
input_xlsx	a list of strings, or NULL. Each string is the absolute or relative path to one dataset's full Excel file, in text format. Must be a list.
input_type	one of c("txt","xlsx"). The format of the input files used. IMPORTANT: if txt, then both input_txt_factors and input_txt_flora must be filled, and have the same length. If xlsx, only input_xlsx should be filled.
concatenate_filters	logical with default to TRUE. See Details below for more informations.

Details

The function will ask questions to the user. When it asks if it should use a filter, the answer should simply be either y (yes) or n (no). If the answer is yes, the function will ask for which values of the given filter to keep, after displaying all availaible values, paired with a number. Answer should be the numbers corresponding to the wanted values.

The function can be used in two "modes", depending on weither concatenate_filters is TRUE or FALSE. If TRUE, then the function will return a single matrix corresponding to all available data for chosen filters, but concatenated. Thus, chosing a filter and setting it to "all" is the same as not chosing the filter. In contrast, if the parameter is set to FALSE, the function will return one matrix per filter combinations (provided that there are data available). For instance, let the filters be : "pluvial", "Cameroun", "Côte d'Ivoire" and "canne" (and no climate filter). If concatenate_filters is TRUE, then the funtion will return one matrix containing all data for sugarcane in rainfed crops, in Cameroun AND Côte d'Ivoire. If FALSE, then two matrix will be return : one for rainfed sugarcane in Cameroun, and one for rainfed sugarcane in Côte d'Ivoire.

Value

A list of matrix. Each matrix contains the concatenated abundance matrix for the chosen factor combination. This list of matrix is ready to be used throught analysis functions.

Examples

```
## Not run:
## Factors
input_files_factors <- list("../Data/DonneesFinales/BEN-MAR-2013-DIV-AD/BEN-MAR-2013-DIV-FAC.txt",
                  "../Data/DonneesFinales/CAM-LEB-1990-DIV-AD/CAM-LEB-1990-DIV-FAC.txt",
```

```
                    "../Data/DonneesFinales/CAM-MAR-1999-MSK-AD/CAM-MAR-1999-MSK-FAC.txt",
                    "../Data/DonneesFinales/CAM-VAL-2001-COT-RE/CAM-VAL-2001-COT-FAC.txt",
                    "../Data/DonneesFinales/CDI-AWA-2011-VIV-AD/CDI-AWA-2011-VIV-FAC.txt",
                    "../Data/DonneesFinales/CDI-GNA-1997-JAC-AD/CDI-GNA-1997-JAC-FAC.txt"
)

## Flora
input_files_flora <- list("../Data/DonneesFinales/BEN-MAR-2013-DIV-AD/BEN-MAR-2013-DIV-AD-FLO.txt",
                  "../Data/DonneesFinales/CAM-LEB-1990-DIV-AD/CAM-LEB-1990-DIV-AD-FLO.txt",
                  "../Data/DonneesFinales/CAM-MAR-1999-MSK-AD/CAM-MAR-1999-MSK-AD-FLO.txt",
                  "../Data/DonneesFinales/CAM-VAL-2001-COT-RE/CAM-VAL-2001-COT-AD-FLO.txt",
                  "../Data/DonneesFinales/CDI-AWA-2011-VIV-AD/CDI-AWA-2011-VIV-AD-FLO.txt",
                  "../Data/DonneesFinales/CDI-GNA-1997-JAC-AD/CDI-GNA-1997-JAC-AD-FLO.txt"
)

filtered_data <- filter_datasets(input_files_factors, input_files_flora, input_type = "txt")

## End(Not run)
```

`get_metrics` *Extract prediction metrics*

Description

This function can be used to get goodness of prediction indicators, from previously trained models.

Usage

```
get_metrics(models, species = "all")
```

Arguments

`models`	a list of random and probability forests. This parameter must match the returned value of `train_models`.
`species`	a character string vector, one (or more) of the species the model was trained with, or `"all"`. The list of every species known by the models can be obtained with `names(models)` (assuming the models were stored in a variable named `model`).

Value

Prints metrics informations, but do not return anything.

Examples

```
## Not run:
factors <- read.csv("Amatrop/var.csv", row.names = 1, stringsAsFactors = TRUE)
AD <- read.csv("Amatrop/ab.csv", row.names = 1)

finalModels <- load_models("./Amatrop/Modeles/randomForests.rds")

get_metrics(finalModels, species = "all")

## End(Not run)
```

`infestation_diagram` *Plot infestation diagrams*

Description

This function can be used to plot and save the infestation diagrams of weed species, for any given list of datasets. plots are saved according to `save_folder` parameter, in `pdf` format.

Usage

```
infestation_diagram(
  data,
  save_folder,
  rf_threshold = NULL,
  am_threshold = NULL,
  main = "auto"
)
```

Arguments

`data`	a list of matrix or data.frames. The (filtered or not) datasets to run analysis on.
`save_folder`	a character string, giving the path to an existing folder, in which all results will be saved.
`rf_threshold`	a single numeric value, corresponding to the treshold value for relative frequency. If a species is less frequent than the threshold, its representation on the plot will only be a point. If it is more frequent, then the EPPO code is displayed. Default to `NULL`, which means that a default value will be given.
`am_threshold`	a single numeric value, corresponding to the treshold value for mean abundance. If a species is less abundant than the threshold, its representation on the plot will only be a point. If it is more abundant, then the EPPO code is displayed. Default to `NULL`, which means that t a default value will be given.
`main`	a character string giving the title of the plot(s). If "auto", then plot(s) will be names automatically.

Value

The function does not return anything. Instead, plots are saved in user's computer.

Examples

```
## Not run:
input_files_factors <- list("../Data/DonneesFinales/BEN-MAR-2013-DIV-AD/BEN-MAR-2013-DIV-FAC.txt",
                   "../Data/DonneesFinales/CAM-LEB-1990-DIV-AD/CAM-LEB-1990-DIV-FAC.txt",
                   "../Data/DonneesFinales/CAM-MAR-1999-MSK-AD/CAM-MAR-1999-MSK-FAC.txt",
                   "../Data/DonneesFinales/CAM-VAL-2001-COT-RE/CAM-VAL-2001-COT-FAC.txt",
                   "../Data/DonneesFinales/CDI-AWA-2011-VIV-AD/CDI-AWA-2011-VIV-FAC.txt",
                   "../Data/DonneesFinales/CDI-GNA-1997-JAC-AD/CDI-GNA-1997-JAC-FAC.txt"
)

input_files_flora <- list("../Data/DonneesFinales/BEN-MAR-2013-DIV-AD/BEN-MAR-2013-DIV-AD-FLO.txt",
                  "../Data/DonneesFinales/CAM-LEB-1990-DIV-AD/CAM-LEB-1990-DIV-AD-FLO.txt",
```

```
                    "../Data/DonneesFinales/CAM-MAR-1999-MSK-AD/CAM-MAR-1999-MSK-AD-FLO.txt",
                    "../Data/DonneesFinales/CAM-VAL-2001-COT-RE/CAM-VAL-2001-COT-AD-FLO.txt",
                    "../Data/DonneesFinales/CDI-AWA-2011-VIV-AD/CDI-AWA-2011-VIV-AD-FLO.txt",
                    "../Data/DonneesFinales/CDI-GNA-1997-JAC-AD/CDI-GNA-1997-JAC-AD-FLO.txt"
)

filtered_data <- filter_datasets(input_files_factors, input_files_flora, input_type = "txt")

infestation_diagram(filtered_data, save_folder = "../Data/infest_diag/")

## End(Not run)
```

`load_models` *Load random forests models from the Amatrop project*

Description

This function can be used to load randome forests models from the Amatrop project. The models are stored in a `.rds` file, and are predictive models of the presence/absence of flora species.

Usage

```
load_models(model_path)
```

Arguments

`model_path`	a character string giving the path to the models location

Value

Returns a list of random forest models. Each element of the returned value match a single species, and is a list of two : a classic random forest model, and a probability forest. Both are object of class `ranger`.

Examples

```
## Not run:
finalModels <- load_models("./Amatrop/Modeles/randomForests.rds")

## End(Not run)
```

PCAIV *Performs Principal Component Analysis with respect to Instrumental Variables*

Description

This function can be used to perform a PCAIV, or Principal Component Analysis with respect to Instrumental Variables, on a set of species abundance and a set of descriptors, within the Amatrop project. The descriptors are supposed to influence the structure of the species abundances. This is mainly a wrapper around ade4's pcaiv function.

Usage

```
PCAIV(
  species,
  factors,
  freq_thresh = 0.1,
  plot = TRUE,
  signif = 0.05,
  scale = FALSE,
  center = FALSE,
  use = "classic",
  adjustNames = FALSE
)
```

Arguments

species	a matrix or data.frame containing the abundance values per species. Each column should correspond to a single species.
factors	a matrix or data.frame containing the descriptors. Each column should correspond to a descriptor. Descriptors can be both numerical or categorical.
freq_thresh	the (relative) frequency threshold. Species less frequent than the threshold will be considered as noise and removed. Default to 0.1.
plot	logical indicating wether to plot the results or not. If TRUE (default), three plots will be displayed in new dev windows : covariance between species and the factorial axes, correlation between species and the factorial axes, and correlation between descriptors and the factorial axes. For correlation plots, only the correlation having p-values smaller than a significance level will be displayed.
signif	the significance level for a species/descriptor correlation to be displayed on plots. Default to 0.05.
scale	a logical value indicating whether the column of the species matrix should be normed.
center	a logical or numeric value, centring option for the species matrix. If TRUE, centring by the mean; if FALSE no centring; if a numeric vector, its length must be equal to the number of columns of the species matrix and give the decentring.
use	one of "classic" or "ortho". It will only change slightly what is printed by the function.

adjustNames — logical value indicating if the column names and terms should be checked and automatically adjusted if necessary. This can be usefull in case there are somme special character in the factors matrix that are incompatible with formulas, but will remove those characters, and thus change ever so slightly the name/term spelling in the results. Consider turning this to TRUE if you enconter something like Error in str2lang(x) : <text>:1:389: unexpected symbol.

Value

Returns a invisible copy of an object of class pcaiv.

Examples

```
## Not run:
factors <- read.csv("Amatrop/var.csv", row.names = 1, stringsAsFactors = TRUE)
AD <- read.csv("Amatrop/ab.csv", row.names = 1)

PCAIV(AD, factors)

## End(Not run)
```

PCAIV_ortho — *Performs orthogonal Principal Component Analysis with respect to Instrumental Variables*

Description

This function can be used to perform an orthogonal PCAIV, or Principal Component Analysis with respect to Instrumental Variables, on a set of species abundance and a set of descriptors, within the Amatrop project. The function will ask which descriptor(s) the PCAIV should be orthogonal to. This is mainly a wrapper around ade4's pcaiv function.

Usage

```
PCAIV_ortho(
  species,
  factors,
  freq_thresh = 0.1,
  plot = TRUE,
  signif = 0.05,
  scale = FALSE,
  center = FALSE,
  adjustNames = FALSE
)
```

Arguments

species — a matrix or data.frame containing the abundance values per species. Each column should correspond to a single species.

factors — a matrix or data.frame containing the descriptors. Each column should correspond to a descriptor. Descriptors can be both numerical or categorical.

freq_thresh	the (relative) frequency threshold. Species less frequent than the threshold will be considered as noise and removed. Default to 0.1.
plot	logical indicating wether to plot the results or not. If TRUE (default), three plots will be displayed in new dev windows : covariance between species and the factorial axes, correlation between species and the factorial axes, and correlation between descriptors and the factorial axes. For correlation plots, only the correlation having p-values smaller than a significance level will be displayed.
signif	the significance level for a species/descriptor correlation to be displayed on plots. Default to 0.05.
scale	a logical value indicating whether the column of the species matrix should be normed.
center	a logical or numeric value, centring option for the species matrix. If TRUE, centring by the mean; if FALSE no centring; if a numeric vector, its length must be equal to the number of columns of the species matrix and give the decentring.
adjustNames	logical value indicating if the column names and terms should be checked and automatically adjusted if necessary. This can be usefull in case there are somme special character in the factors matrix that are incompatible with formulas, but will remove those characters, and thus change ever so slightly the name/term spelling in the results. consider turning this to TRUE if you enconter something like Error in str2lang(x) : <text>:1:389: unexpected symbol

Value

Returns a invisible copy of an object of class pcaiv.

Examples

```
## Not run:
factors <- read.csv("Amatrop/var.csv", row.names = 1, stringsAsFactors = TRUE)
AD <- read.csv("Amatrop/ab.csv", row.names = 1)

PCAIV_ortho(AD, factors)

## End(Not run)
```

---

plot_corr_cooc *Plot correlation and co-occurence matrix*

Description

This function can be used to plot and save the correlation and co-occurence of weed species, for any given list of datasets. plots are saved according to save_folder parameter, in .pdf format.

Usage

```
plot_corr_cooc(data, save_folder, n_species = 50, title = "auto", ...)
```

Arguments

data	a list of matrix or data.frames. The (filtered or not) datasets to run analysis on.
save_folder	a character string, giving the path to an existing folder, in which all results will be saved.
n_species	a single numeric value, corresponding to the number of species to include in the plot. Species are sorted by relative frequency, meaning that `n_species = 50` will display the 50 more frequent species. Default to 50. If a dataset contains less than `n_sepecies` species, all species will be displayed.
title	a vector of two character strings giving the title of the each of the plots, or `"auto"`. If `"auto"`, then plots will be names automatically.
...	additionnal graphic parameters to be passed to the plotting routines.

Value

The function does not return anything. Instead, plots are saved in user's computer.

Examples

```
## Not run:
input_files_factors <- list("../Data/DonneesFinales/BEN-MAR-2013-DIV-AD/BEN-MAR-2013-DIV-FAC.txt",
                   "../Data/DonneesFinales/CAM-LEB-1990-DIV-AD/CAM-LEB-1990-DIV-FAC.txt",
                   "../Data/DonneesFinales/CAM-MAR-1999-MSK-AD/CAM-MAR-1999-MSK-FAC.txt",
                   "../Data/DonneesFinales/CAM-VAL-2001-COT-RE/CAM-VAL-2001-COT-FAC.txt",
                   "../Data/DonneesFinales/CDI-AWA-2011-VIV-AD/CDI-AWA-2011-VIV-FAC.txt",
                   "../Data/DonneesFinales/CDI-GNA-1997-JAC-AD/CDI-GNA-1997-JAC-FAC.txt"
)

input_files_flora <- list("../Data/DonneesFinales/BEN-MAR-2013-DIV-AD/BEN-MAR-2013-DIV-AD-FLO.txt",
                  "../Data/DonneesFinales/CAM-LEB-1990-DIV-AD/CAM-LEB-1990-DIV-AD-FLO.txt",
                  "../Data/DonneesFinales/CAM-MAR-1999-MSK-AD/CAM-MAR-1999-MSK-AD-FLO.txt",
                  "../Data/DonneesFinales/CAM-VAL-2001-COT-RE/CAM-VAL-2001-COT-AD-FLO.txt",
                  "../Data/DonneesFinales/CDI-AWA-2011-VIV-AD/CDI-AWA-2011-VIV-AD-FLO.txt",
                  "../Data/DonneesFinales/CDI-GNA-1997-JAC-AD/CDI-GNA-1997-JAC-AD-FLO.txt"
)

filtered_data <- filter_datasets(input_files_factors, input_files_flora, input_type = "txt")

plot_corr_cooc(filtered_data, save_folder = "../Data/corr_cooc/")

## End(Not run)
```

---

predict_species *Predict species presence/absence*

Description

This function can be used to predict the presence (or probability of presence) for one or more species, using the previously trained models and a new set of observations for the descriptors.

Usage

```
predict_species(models, newdata, species = "all", type = "proba")
```

Arguments

models	a list of random and probability forests. This parameter must match the returned value of train_models.
newdata	a matrix or data.frame containing new observations for the descriptors. Each column should correspond to a descriptor, and the descriptors names must match exactly the ones the model was trained with. Each row is a new observation. Descriptors can be both numerical or categorical.
species	a character string vector, one (or more) of the species the model was trained with, or "all". The list of every species known by the models can be obtained with names(models) (assuming the models were stored in a variable named models).
type	one of "presence","proba", with default to the latter. "presence" will give the prediction in presence/absence, and "proba" will give the predicted probability of presence.

Value

Returns a invisible copy of a data.frame containing the predictions. These are also printed.

Examples

```
## Not run:
factors <- read.csv("Amatrop/var.csv", row.names = 1, stringsAsFactors = TRUE)
AD <- read.csv("Amatrop/ab.csv", row.names = 1)

finalModels <- load_models("./Amatrop/Modeles/randomForests.rds")

n <- 10
newdata <- data.frame(alt = sample(factors$alt, n, replace = TRUE),
                      cult = sample(factors$cult, n, replace = TRUE),
                      desherb = sample(factors$desherb, n, replace = TRUE),
                      irrig = sample(factors$irrig, n, replace = TRUE),
                      pluvio = sample(factors$pluvio, n, replace = TRUE),
                      preced = sample(factors$preced, n, replace = TRUE)
)
rownames(newdata) <- paste("R", 1:n, sep = "")
predict_species(finalModels, newdata, type = "proba", species = "all")

## End(Not run)
```

---

relat_freq *Compute relative frequency of species*

Description

This function can be used to compute the relative frequency of weed species accross all flora readings, for given list of datasets.

Usage

```
relat_freq(data)
```

Arguments

data	a list of data.frames of matrix. Each data.frame or matrix is one dataset, filtered or not.

Value

a list of named numeric vectors, each corresponding to one of the entrie datasets.

---

save_analysis *Save basic analysis on datasets*

Description

This function can be used to perform and save the results of frequency and abundance analysis. This function computes the relative frequency and mean abundance of species accross flora readings of given datasets, and save the results in a csv file, which location will be according to the 'save_folder parameter. User can choose to add the species reference trought 'ref' parameter, which will add more informations on species in the results.

Usage

```
save_analysis(data, save_folder, ref = NULL)
```

Arguments

data	a list of matrix or data.frames. The (filtered or not) datasets to run analysis on.
save_folder	a character string, giving the path to an existing folder, in which all results will be saved.
ref	a character string, giving the path to the flora reference Excel file, or NULL. If given, then the results will contain additionnal columns corresponding to taxa, family name, and most used synonyms. If not provided, the only available information on species in the results will be their EPPO codes.

Value

The function does not return anything. Instead, results are saved in user's computer.

Examples

```
## Not run:
input_files_factors <- list("../Data/DonneesFinales/BEN-MAR-2013-DIV-AD/BEN-MAR-2013-DIV-FAC.txt",
                  "../Data/DonneesFinales/CAM-LEB-1990-DIV-AD/CAM-LEB-1990-DIV-FAC.txt",
                  "../Data/DonneesFinales/CAM-MAR-1999-MSK-AD/CAM-MAR-1999-MSK-FAC.txt",
                  "../Data/DonneesFinales/CAM-VAL-2001-COT-RE/CAM-VAL-2001-COT-FAC.txt",
                  "../Data/DonneesFinales/CDI-AWA-2011-VIV-AD/CDI-AWA-2011-VIV-FAC.txt",
                  "../Data/DonneesFinales/CDI-GNA-1997-JAC-AD/CDI-GNA-1997-JAC-FAC.txt"
)
```

```
input_files_flora <- list("../Data/DonneesFinales/BEN-MAR-2013-DIV-AD/BEN-MAR-2013-DIV-AD-FLO.txt",
                  "../Data/DonneesFinales/CAM-LEB-1990-DIV-AD/CAM-LEB-1990-DIV-AD-FLO.txt",
                  "../Data/DonneesFinales/CAM-MAR-1999-MSK-AD/CAM-MAR-1999-MSK-AD-FLO.txt",
                  "../Data/DonneesFinales/CAM-VAL-2001-COT-RE/CAM-VAL-2001-COT-AD-FLO.txt",
                  "../Data/DonneesFinales/CDI-AWA-2011-VIV-AD/CDI-AWA-2011-VIV-AD-FLO.txt",
                  "../Data/DonneesFinales/CDI-GNA-1997-JAC-AD/CDI-GNA-1997-JAC-AD-FLO.txt"
)

filtered_data <- filter_datasets(input_files_factors, input_files_flora, input_type = "txt")

save_analysis(filtered_data, save_folder = "../Data/Results/", ref = "../Data/reference.xlsx")

## End(Not run)
```

`train_models` *Train random and probability forests*

Description

This function can be used to train a set of random and probability forest in order to models, and then predict, species presence/absence, given a set of predictors. It is mainly a wrapper around `caret`'s `train` and the `ranger` functions.

Usage

```
train_models(
  species,
  factors,
  save_path,
  freq_thresh = 0.25,
  trace = TRUE,
  search = "grid",
 tunegrid = expand.grid(mtry = 1:4, splitrule = "gini", min.node.size = c(10, 20))
)
```

Arguments

`species`	a matrix or data.frame containing the presence/absence values per species. Presence/absence should be described by 1 and 0 respectively. Each column should correspond to a single species.
`factors`	a matrix or data.frame containing the descriptors. Each column should correspond to a descriptor. Descriptors can be both numerical or categorical.
`save_path`	a character string giving a valid path to the location the models should be stored. If `NULL`, the models are not saved on the computer.
`freq_thresh`	the (relative) frequency threshold. Species less frequent than the threshold will be considered as not relevant and removed. Default to `0.25`.
`trace`	logical indicating wether the function should display informations on the training's progression.
`search`	one of `"random"`, `"grid"`, with default to the latter. Describe how the tuning parameter grid is determined during hyperparameter optimisation. In most cases, small hyperparameter spaces work best with `"grid"`, and vice versa.

tunegrid — a data.frame giving the hyperparameter spaces over which the hyperparameter optimisaton will be done. Each line of the data.frame should match a single hyperparameters combination.

Value

Returns a invisible copy of a list of random forest models. Each element of the returned value match a single species, and is a list of two : a classic random forest model, and a probability forest. Both are object of class ranger.

Examples

```
## Not run:
factors <- read.csv("Amatrop/var.csv", row.names = 1, stringsAsFactors = TRUE)
AD <- read.csv("Amatrop/ab.csv", row.names = 1)

models <- train_models(AD, factors, save_path = "./Amatrop/Modeles/randomForests.rds")

## End(Not run)
```

---

xlsx_to_txt *Generate text format from base original dataset*

Description

This function can be used to create text formated version of spreadsheets of the original datasets from the Amatrop project. Original datasets should be Excel files (with .xlsx extension). The function creates the text formated files, and store them in the user's computer according to output_path parameter. Three kinds of files can be created, corresponding to the three possible kinds of spreadsheets : factors array, abundance flora array, or presence/absence flora array. The function can create files for one, two, or all spreadsheets at once.

Usage

```
xlsx_to_txt(files, output_path = "auto", kind = "all")
```

Arguments

files — a list of strings. Each string is the absolute or relative path to one dataset. Must be a list.

output_path — a character string vector of the same size than files giving the path to the folders to store each text file created, or "auto". If "auto", the function will try to detect the path automatically from the input path. Default to "auto".

kind — a character string vector. The type of array created, corresponding to spreadsheets in the original datasets. Should be one or more of c("abundance","presence-absence","fa
If "all", then all three text files will be created. Default to "all".

Value

The function does not return anything.

Références

[1] Creative commons licence, attribution 4.0 international (cc by 4.0), 2020. https://creativecommons.org/licenses/by/4.0/.

[2] N. Abramson. *Information theory and coding*. MC Graw Hill, 1963.

[3] S. Baillif, J. Esther, P. Marnotte, D. Marion, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Effect of tillage date on germination and phenology of weeds in sugarcane, at etang salé, reunion island (2017-2018), 2020. doi:10.18167/DVN1/GDNUCH, CIRAD Dataverse, V1.

[4] M. Boraud, P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed surveys of several sugarcane estates in côte d'ivoire (1997), 2020. doi:10.18167/DVN1/Y5SRMR, CIRAD Dataverse, V2.

[5] J. (Josias) Braun-Blanquet, Henry Shoemaker Conard, and George D. Fuller. *Plant sociology: the study of plant communities*. New York and London, McGraw-Hill book company, inc., 1932. https://www.biodiversitylibrary.org/bibliography/7161.

[6] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001.

[7] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

[8] Merce Crosas. The dataverse network: An open-source application for sharing, discovering and preserving data. *D-Lib Magazine*, Volume 17, 2011. http://www.dlib.org/dlib/january11/crosas/01crosas.html.

[9] Ph. Daget, M. Godron, J. L. Guillerm, I. Drdos, H. Ruzickova, and E. Urvichiarova. *Profils Écologiques et Information Mutuelle Entre Especes et Facteurs Écologiques*, pages 121–149. Springer Netherlands, Dordrecht, 1972. https://doi.org/10.1007/978-94-015-7241-5_10.

[10] EPPO, 2020. EPPO Global Database (available online). https://gd.eppo.int. Retrieved January to April 2020.

[11] R. A. Fisher. On the interpretation of X^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922. http://www.jstor.org/stable/2340521.

[12] David Freedman, Robert Pisani, and Roger Purves. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[13] G.M. Gnahoua, P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed survey in fallows in oumé region, côte d'ivoire (1997), 2020. doi:10.18167/DVN1/RRQWXJ, CIRAD Dataverse, V2.

[14] Michel Godron. *Application de la Theorie de L'information a L'etude de L'homogeneite et de la Structure de la Vegetation*, pages 31–38. Springer Netherlands, Dordrecht, 1970. https://doi.org/10.1007/978-94-010-3353-4_4.

[15] S. Huguenot, Aboubakary, E. Vall, P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Effect of mechanical weeding compared to hand weeding on cotton weed flora in northern cameroon (2001), 2020. doi:10.18167/DVN1/9PDXYC, CIRAD Dataverse, V2.

[16] J. Ipou Ipou, P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Synthesis of the inventory of weeds of food crops in côte d'ivoire (2009-2016), 2020. , doi:10.18167/DVN1/V0QGKT, CIRAD Dataverse, V2.

[17] S.A. Kouamé, J. Ipou Ipou, P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed survey in food crops in adbijan disctrict, côte d'ivoire (2015), 2020. doi:10.18167/DVN1/VCLIB5, CIRAD Dataverse, V2.

[18] T. Le Bourgeois, A. Berton, P. Marnotte, S. Auzoux, and B. Fayolle. Weed survey in vegetables and orchards in french guiana (2018), 2020. doi:10.18167/DVN1/TMZSMW, CIRAD Dataverse, V2.

[19] T. Le BOURGEOIS and J. L. GUILLERM. Etendue de distribution et degré d'infestation des adventices dans la rotation cotonnière au nord-cameroun. *Weed Research*, 35(2):89–98, 1995. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3180.1995.tb02021.x.

[20] T. Le Bourgeois, P. Marnotte, S. Auzoux, and B. Fayolle. Weed survey in cotton rotation in northern cameroon (1988-1992), 2020. doi:10.18167/DVN1/5ZEQGO, CIRAD Dataverse, V2.

[21] T. Le Bourgeois, P. Marnotte, S. Auzoux, and B. Fayolle. Weed survey in various cropping systems in the reunion island (2003-2006), 2020. doi:10.18167/DVN1/UAHMEJ, CIRAD Dataverse, V2.

[22] Thomas Le Bourgeois, Alain Paul Andrianaivo, Pierre Grard, Azaad Gaungoo, Ibrahim Yahaya, Jean Augustin Randriamampianina, Balasubramanian Dhandapani, Pascal Marnotte, Vincent Blanfort, Prabhakar Rajagopal, Thomas Vattakaven, and Choukry Kazi Tani. Wiktrop - weed identification and knowledge in the tropics and mediterranean area - web 2.0 participatory portal, 2019. European Union programme ACP S&T II, Cirad, IFP, MCIA/MSIRI, FOFIFA, CNDRS eds. http://portal.wiktrop.org.

[23] Thomas Le Bourgeois, A. Berton, Vincent Blanfort, Azaad Gaungoo, Pierre Grard, Pascal Marnotte, Prabhakar Rajagopal, Jean Augustin Randriamampianina, Thomas Vattakaven, and Ibrahim Yahaya. Enjeux et contraintes du partage et de la diffusion des connaissances en malherbologie tropicale pour une meilleure gestion des enherbements, exemple du portail collaboratif wiktrop. In *24e Conférence du COLUMA : Journées internationales sur la lutte contre les mauvaises herbes*, 2019. Végéphyl. Alfortville : Végéphyl, 9 p. Conférence du COLUMA : Journées internationales sur la lutte contre les mauvaises herbes. 24, Orléans, France, 3 Décembre 2019/5 Décembre 2019. https://agritrop.cirad.fr/594419/.

[24] Thomas Le Bourgeois, Vincent V. Blanfort, Cédric Péret, Vincent Petiot, and Jean-Marie Capron. The WIKTROP collaborative portal : sharing and disseminating knowledge on weeds of tropical pastures for better management of grazing land degradation. 9th Multi-stakeholder Partnership Meeting – Innovation for Sustainable Livestock Systems, Sep 2019. https://hal.umontpellier.fr/hal-02299209.

[25] Jean-Dominique Lebreton, Robert Sabatier, G. Banco, and A. Bacou. *Principal Component and Correspondence Analyses with Respect to Instrumental Variables : An Overview of Their Role in Studies of Structure - Activity and Species - Environment Relationships*, pages 85–114. 01 1991.

[26] James Malley, J Kruppa, Abhijit Dasgupta, Karen Malley, and Andreas Ziegler. Probability machines consistent probability estimation using nonparametric learning machines. *Methods of information in medicine*, 51:74–81, 09 2011.

[27] P. Marnotte, Jean Jo Esther, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed survey in herbicide experiments in sugarcane in reunion island (2005-2016), 2020. doi:10.18167/DVN1/OFZEH1, CIRAD Dataverse, V2.

[28] P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Effect of tillage date on germination and phenology of weeds in sugarcane, at bassin plat, reunion island (2019-2020), 2020. doi:10.18167/DVN1/1HCKBU, CIRAD Dataverse, V1.

[29] P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed survey in cotton crop in guinea (1996), 2020. doi:10.18167/DVN1/EVCQFR, CIRAD Dataverse, V2.

[30] P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed survey in sugarcane in zuenoula, côte d'ivoire (1991-1992), 2020. doi:10.18167/DVN1/3KAVBK, CIRAD Dataverse, V2.

[31] P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed survey in vegetable growing in mauritius (2019), 2020. doi:10.18167/DVN1/VBE4VT, CIRAD Dataverse, V2.

[32] P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed surveys in food crops and irrigated rice in the zonmon region, benin (2013), 2020. doi:10.18167/DVN1/X98KVY, CIRAD Dataverse, V1.

[33] P. Marnotte, D. Marion, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Effect of tillage date on germination and phenology of weeds in sugarcane, at la mare, reunion island (2019-2020), 2020. doi:10.18167/DVN1/FWFAUY, CIRAD Dataverse, V1.

[34] P. Marnotte, B. Mathieu, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed surveys in muskwari sorghum in northern cameroon (1999), 2020. doi:10.18167/DVN1/H7AJUK, CIRAD Dataverse, V2.

[35] Pascal Marnotte. Influence des facteurs agroécologiques sur le développement des mauvaises herbes en climat tropical humide, 1984. Biol.et Syst. des Mauvaises Herbes. Paris, France: COLUMA-EWRS.

[36] Pascal Marnotte and Mathieu Bertrand. L'enherbement des sols à muskuwaari au nord-cameroun. In *Onzième colloque international sur la biologie des mauvaises herbes*, 2000. AFPP, INRA. Paris : AFPP, 151-158. (AFPP Annales) ISBN 2-905550-87-2 Colloque International sur la Biologie des Mauvaises Herbes. 11, Dijon, France, 6 Septembre 2000/8 Septembre 2000. http://agritrop.cirad.fr/476748/.

[37] POWO, 2019. Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Published on the Internet; http://www.plantsoftheworldonline.org/ Retrieved January to April 2020.

[38] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. https://www.R-project.org.

[39] A. Rafenomanjato, A. Ripoche, P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed survey in upland rice in madagascar (2017), 2020. doi:10.18167/DVN1/VSY7Y5, CIRAD Dataverse, V3.

[40] J.A. Randriamampianina, P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed survey in various cropping systems in madagascar (1998), 2020. doi:10.18167/DVN1/O1EHV2, CIRAD Dataverse, V2.

[41] J.A. Randriamampianina, P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed survey in various cropping systems in madagascar (1999), 2020. doi:10.18167/DVN1/XZ4CBA, CIRAD Dataverse, V2.

[42] F. Rokotonirina, P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed survey in upland rice cropping system in madagascar (2015), 2020. doi:10.18167/DVN1/NQYPEY, CIRAD Dataverse, V2.

[43] Y. Roskov, G. Ower, T. Orrell, D. Nicolson, N. Bailly, P.M. Kirk, T. Bourgoin, R.E. DeWalt, W. Decock, E. van Nieukerken, J. Zarucchi, and Penev L., 2019. Species 2000 & ITIS Catalogue of Life, 2019 Annual Checklist. Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-884X. www.catalogueoflife.org/annual-checklist/2019.

[44] Robert Sabatier, Jean-Dominique Lebreton, and Chessel Daniel. Principal component analysis with instrumental variables as a tool for modelling composition data. *Multiway data analysis*, pages 341–352, 10 1989.

[45] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471, 1987.

[46] V. Stevoux, P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed survey in upland rice in northern viet nam (1999), 2020. doi:10.18167/DVN1/MDP0H5, CIRAD Dataverse, V2.

[47] K.E. Téhia, P. Marnotte, Jean Jo Esther, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Survey of major weeds of cotton crop in côte d'ivoire (2014), 2020. doi:10.18167/DVN1/TCCSRC, CIRAD Dataverse, V2.

[48] A. Touré, J. Ipou Ipou, P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Weed surveys in food crops in center-east côte d'ivoire (2011-2014), 2020. doi:10.18167/DVN1/ZH4W5M, CIRAD Dataverse, V2.

[49] P. Viaud, M. Christina, P. Marnotte, T. Le Bourgeois, S. Auzoux, and B. Fayolle. Effect of cover crop, irrigation and fertilizer on weeds in sugarcane in reunion island (2019-2021), 2020. doi:10.18167/DVN1/YWNUCG, CIRAD Dataverse, V1.